Manufacturing Engineer's Reference Book

Edited by
Dal Koshal
with specialist contributors

BUTTERWORTH
HEINEMANN

Butterworth-Heinemann Ltd
Linacre House, Jordan Hill, Oxford OX2 8DP

A member of the Reed Elsevier group

OXFORD LONDON BOSTON
MUNICH NEW DELHI SINGAPORE SYDNEY
TOKYO TORONTO WELLINGTON

First published 1993

© Butterworth-Heinemann 1993

All rights reserved. No part of this publication may be reproduced in any material form (including photocopying or storing in any medium by electronic means and whether or not transiently or incidentally to some other use of this publication) without the written permission of the copyright holder except in accordance with the provisions of the Copyright, Designs and Patents Act 1988 or under the terms of a licence issued by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London, England W1P 9HE. Applications for the copyright holder's written permission to reproduce any part of this publication should be addressed to the publishers

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN 0 7506 1154 5

Library of Congress Cataloguing in Publication Data
A catalogue record for this book is available from the Library of Congress

Every effort has been made to trace the holders of copyright material. However, if any omissions have been made, the authors will be pleased to rectify them in future editions of the book.

Printed and bound in Great Britain by Bath Press Ltd, Avon

List of Contributors

T Z Blazynski PhD, BSc(Eng), CEng, MIMechE
Formerly Reader in Applied Plasticity
Department of Mechanical Engineering
University of Leeds

John Brydson
Former Head of Department of Physical Sciences and Technology
Polytechnic of North London

John L Burbidge OBE, DSc, HonFIProdE, FIMechE, FBIM
Visiting Professor in Manufacturing Systems
Cranfield Institute of Technology

T C Buttery BSc, PhD, CEng, MIEE, AMIM
Senior Research Co-ordinator at CIMTEX
School of Engineering and Manufacture
De Montfort University, Leicester

Harry L Cather CEng, MIEE, MBIM, MBA, MSc
Senior Lecturer
Department of Mechanical and Manufacturing Engineering
University of Brighton

E N Corlett DSc, PhD, MEng, BSc(Eng), FEng, FIMechE, FIEE, HonFErgS, CPsychol
Emeritus Professor
Institute of Occupational Ergonomics
University of Nottingham
Nottingham NG7 2RD

Roy D Cullum FIED
Editor, *Materials and Manufacture*
Publication Services, Worthing

Brad D Etter PhD
Assistant Professor, Bioengineering Program
Texas A&M University, USA

William T File CEng, MIMechE
William T File & Associates
Consultants in Maintenance Engineering
Aylesbury, Bucks

C J Fraser BSc, PhD, CEng, FIMechE, MInstPet
Reader in Mechanical Engineering
Department of Mechanical Engineering
Dundee Institute of Technology

Jan Glownia PhD, Eng
Academy of Mining and Metallurgy
Institute of Foundry Technology and Mechanisation
Cracow, Poland

R Goss BEng, CEng, MIMechE
Senior Application System Engineer (Industrial Product Division)
Loctite (UK) Ltd

E N Gregory CEng, FIM, FWeldI
Head of Advisory Section
The Welding Institute
Cambridge

John Hunt CEng, MIEE
Director
Custom Engineering and Networks
Bookham, Surrey

D Koshal MSc, PhD, CEng, FIMechE, FIEE
Principal Lecturer, Manufacturing Engineering
Department of Mechanical and Manufacturing Engineering
University of Brighton

Gordon M Mair BSc(Hons), DMS, CEng, MIEE. MBIM
Lecturer
Department of Design, Manufacture and Engineering Management (MEM Division)
University of Strathclyde

Gerald E Miller PhD
Professor and Chairman, Bioengineering Program
Texas A&M University, USA

John S Milne BSc(Hons), CEng, FIMechE
Senior Lecturer in Mechanical Engineering
Department of Mechanical Engineering
Dundee Institute of Technology

D B Richardson MPhil, DIC, CEng, FIMechE, FIEE
Formerly Principal Lecturer in Manufacturing Engineering
Department of Mechanical and Manufacturing Engineering
University of Brighton

Leslie M Wyatt MA(Cantab), FMI, CEng
Independent Consultant and Technical Author

Preface

No reference book on manufacturing engineering will ever be an unqualified success. It cannot be a panacea for those seeking cures for the ills of poor management. At best it provides an insight into the multifarious techniques and processes which combine to enable goods to be produced competitively in terms of cost, quality, reliability and delivery.

Within manufacturing organisations it is to be hoped that specialist knowledge exists in all the relevant fields, extending beyond the confines of the chapters in a reference book. If it does not exist, the companies must ensure that they acquire it by providing suitable training or by employing experts, if they hope to survive.

What purpose is served, then, by this sort of reference book? The editor believes that, apart from those copies which are destined to gather dust in executives' bookcases to help provide an ambience of professional respectability, the book will be useful for top and middle managers who feel the need to widen their perspectives. With this in mind it has been written in compartmentalised form, each section being free-standing and capable of being understood as an introductory text. It will also be of use to engineering students as an adjunct to the more specific texts used in support of their lectures.

Some reference books are primarily compendiums of tabulated data, essential to the task of technical quantification. This volume is not such a compendium: it is mainly a qualitative approach to knowledge gathering from which subsequent quantification may be attempted. It covers those aspects of manufacture which are essential for designing new production systems or for managing exisiting factories. These include materials selection, manufacturing and fabrication processes, quality control, and the use of computers for the control of processes and production management.

The editor wishes to thank the specialist authors who contributed their expertise for their forbearance during the various stages of preparation, together with colleagues, past and present, at the University of Brighton for their suggestions. Special thanks are due to Mr Don Richardson for his helpful suggestions on various chapters.

I would particularly like to thank my wife and family for their continued support during this project.

Dr Dalbir Koshal
University of Brighton

Contents

Preface

List of Contributors

1 Materials Properties and Selection

Engineering properties of materials · The principles underlying materials selection · Ferrous metals · Non-ferrous metals · Composites · Engineering ceramics and glasses

2 Polymers, Plastics and Rubbers

Introduction · General properties of rubber materials · Survey of commerical rubbery materials · General properties of plastics · Survey of commercial plastics materials · Processing of rubbers and plastics · Design of rubber components · Design of plastic components

3 Metal Casting and Moulding Processes

Economics of casting and moulding · Sand casting · Low and high pressure die casting · Investment casting · Shell moudling · Sintering

4 Metal Forming

The origin, nature and utilisation of plastic flow · Process assessment · An outline of the theory of plasticity · Tool design · Rolling processes and products · Forging operations · Extrusion · Cold drawing of wire and tube · Sheet-metal forming · High-energy-rate operations · Superplastic and mashy state forming

5 Large-chip Metal Removal

Large-chip processes · Cutting-tool geometry · Cutting-tool materials · Chip formation and cutting parameters · Forces and power in metal cutting · Surface-finish considerations · Tool-life assessment · Economics of metal cutting

6 Non-chip Metal Removal

Introduction · Electrical processes · Mechanical processes · Thermal processes

7 Electronic Manufacture

Introduction · Assembly of through-hole components · Surface-mount technology · Surface-mount components · Assembly of printed circuit boards · Surface-mount-component attachment methods prior to soldering · Soldering · Cleaning · Automatic testing equipment · Surface-mount-component placement machines · Printed-circuit-board layout · Possible defects during manufacture · Introducing surface-mount technology

8 Metal Finishing Processes

Introduction · Abrasives · Grinding wheels and grinding wheel selection · Mounting the grinding wheel · Balancing and dressing · Grinding mechanics · Wheel wear · Grinding ratio · Grinding forces · Coolant · Grinding processes · Newer abrasives and grinding techniques · Honing · Honing practice · Superfinishing · Coated abrasives · Machining with coated abrasives · Abrasive discs · Lapping · Polishing · Blasting processes

9 Fabrication

Fasteners · Welding, soldering and brazing · Adhesives

10 Electrical and Electronic Principles

Introduction · Basic electrical technology · Analogue and digital electronics theory · Electrical machines · Electrical safety

11 Microprocessors, Instrumentation and Control

Basic control systems · Control strategies · Instrumentation and measurement · Microprocessor technology · Interfacing of computers to systems · Microprocessor-based control · Programmable logic controllers · Robot applications

12 Machine Tool Control Elements

Machine tool control system — overview · Electric motors · The servomotor control and amplifier · Transmission elements · Fluid power actuators · Fluid power actuator control valves · Feedback transducers · Conclusion

13 Communication and Integration Systems

Computer architecture · Computer operating systems · Computer languages · Peripheral devices · Human–computer interfaces · Networks · Databases · Knowledge-based systems

14 Computers in Manufacturing

Introduction · Computer-integrated manufacturing · Manufacturing control systems · Computer-aided design and manufacture · Numerical engineering and control · Flexible manufacturing systems · Industrial robotics and automation

15 Manufacturing and Operations Management

Introduction to manufacturing management · Types of production · Systems theory · Management and the functions · General management · Product design · Production planning · Design of the material flow system · Assembly groups · Marketing · Production control · Purchasing · Finance · Personnel · Secretarial function · Conclusion

16 Manufacturing Strategy

Manufacturing strategy and organisation · Strategies for increasing manufacturing competitiveness

17 Control of Quality

The concept of quality · Quality through integrated design · Standards · Material and process control · Process capability · Product acceptance sampling schemes · Statistical process control · Measurement of form and surface · Non-destructive testing

18 Terotechnology and Maintenance

Management of assets · Life-cycle costing · Plant selection and replacement · Measurement of the effectiveness of maintenance · Operational aspects of maintenance · Preventive maintenance · Condition monitoring · Inventory and maintenance

19 Ergonomics

Introduction · Robotics · Computer-aided manufacture · The automated office · Manual assembly · The economics of ergonomics · Conclusions

1 Materials Properties and Selection

L Wyatt

Contents

1.1 Engineering properties of materials 1/3
1.1.1 Elastic properties 1/3
1.1.2 Tensile testing parameters 1/3
1.1.3 Hardness 1/3
1.1.4 Fracture toughness and impact testing 1/4
1.1.5 Fatigue 1/4
1.1.6 Creep and stress rupture 1/4
1.1.7 Thermal properties 1/5
1.1.8 Electrical properties 1/5
1.1.9 Optical properties 1/5

1.2 The principles underlying materials selection 1/5
1.2.1 Introduction 1/5
1.2.2 Techniques of materials selection 1/6
1.2.3 Preliminary examination of design 1/6
1.2.4 The classical procedure 1/6
1.2.5 Drawbacks of the classical procedure 1/8
1.2.6 The imitative procedure 1/8
1.2.7 The comparative procedure 1/8
1.2.8 Information sources 1/8
1.2.9 Computerisation of materials selection 1/8

1.3 Ferrous metals 1/10
1.3.1 Introduction and standards for wrought steels 1/10
1.3.2 Types of wrought steel 1/10
1.3.3 Steelmaking and casting of wrought-steel ingots 1/12
1.3.4 Mechanical working of metals 1/15
1.3.5 Constitution and heat treatment of ferritic steels 1/21
1.3.6 Carbon/carbon–manganese steels 1/25
1.3.7 High strength low alloy steels 1/28
1.3.8 Electrical steels 1/28
1.3.9 Hardened and tempered steels 1/29
1.3.10 Free cutting steels 1/31
1.3.11 Case hardening steels 1/31
1.3.12 Stainless steels 1/35
1.3.13 Corrosion resistance of stainless steels 1/38
1.3.14 Heat resisting steels 1/43
1.3.15 Toughness in steels 1/44
1.3.16 Maraging steels 1/45
1.3.17 Weldability of steels 1/46
1.3.18 Tool steels 1/48
1.3.19 Steels for springs 1/49
1.3.20 Cast steel 1/50
1.3.21 Cast irons—general 1/50
1.3.22 Grey cast iron 1/50
1.3.23 Nodular graphite iron 1/53
1.3.24 Austempered ductile iron 1/53
1.3.25 Malleable iron 1/54
1.3.26 Austenitic cast irons 1/54
1.3.27 High silicon cast irons 1/54
1.3.28 White cast irons (abrasion resisting white irons) 1/54
1.3.29 High chromium iron 1/54

1.4 Non-ferrous metals 1/54
1.4.1 Copper and its alloys 1/54
1.4.2 Aluminium and its alloys 1/59
1.4.3 Titanium and its alloys 1/66
1.4.4 Magnesium and its alloys 1/74
1.4.5 Nickel and its alloys 1/77
1.4.6 Zinc and its alloys 1/86
1.4.7 Low-melting-point metals, lead, tin and their alloys 1/90
1.4.8 Cobalt and its alloys 1/92
1.4.9 Other non-ferrous metals 1/93

1.5 Composites 1/94
1.5.1 Introduction 1/94
1.5.2 Reinforcing fibres 1/95
1.5.3 Polymer matrices 1/96
1.5.4 Manufacturing procedures for filamentary polymer composites 1/96
1.5.5 Properties of filamentary polymer composites 1/96
1.5.6 Applications of filamentary polymer composites 1/99
1.5.7 Discontinuous fibre reinforced polymer components 1/101
1.5.8 Carbon–carbon composites 1/102
1.5.9 Fibre reinforced metals 1/104
1.5.10 Fibre reinforced glasses and ceramics 1/104
1.5.11 Reinforced concrete 1/106
1.5.12 Particulate composites 1/107
1.5.13 Laminar composites 1/107
1.5.14 Wood and resin-impregnated wood 1/107

1.6 Engineering ceramics and glasses 1/109
- 1.6.1 Introduction 1/109
- 1.6.2 Standards 1/109
- 1.6.3 Clay based ceramics 1/109
- 1.6.4 Oxide based ceramics 1/110
- 1.6.5 Non-oxide ceramics 1/111
- 1.6.6 Carbons and graphites 1/112
- 1.6.7 Miscellaneous ceramics 1/113
- 1.6.8 Glasses 1/113
- 1.6.9 Glass ceramics 1/113
- 1.6.10 Mechanical properties 1/113
- 1.6.11 Manufacturing processes 1/115
- 1.6.12 The future prospects of engineering ceramics 1/115

1.1 Engineering properties of materials

1.1.1 Elastic properties

Elastic or Young's modulus, E (units GPa) The stress required to produce unit strain in the same direction, i, in the absence of restraint in the orthogonal directions:

$$E_i = \sigma_i \varepsilon_i^{-1} \tag{1.1}$$

where σ is the stress and ε the strain which it produces. A standard testing method is described in ASTM E231.

Shear modulus, G (units GPa) The shear stress required to produce unit angular rotation of a line perpendicular to the plane of shear:

$$G = T\phi^{-1} \tag{1.2}$$

where T is the shear stress and ϕ is the angular rotation (in radians).

Bulk modulus, K (units GPa) The hydrostatic pressure p required to effect unit change in volume V:

$$K = pV(\Delta V)^{-1} \tag{1.3}$$

Poisson's ratio ν (dimensionless) The ratio of the strain in a direction orthogonal to the direction of stress to the strain in the direction of stress:

$$\nu = \varepsilon_{j,k} \varepsilon_i^{-1} \tag{1.4}$$

These four basic elastic properties apply to homogeneous and isotropic materials and are related by the equations:

$$E = 3K(1 - 2\nu) \tag{1.5}$$

$$= 2G(1 + \nu) \tag{1.6}$$

In the case of a material which has anisotropic elastic properties the terms used may have different meanings, and stresses and strains should be related using tensor analysis.

1.1.2 Tensile testing parameters

When considering the properties obtained from the tensile test it should be realised that the results are always reported as though the load was applied to the initial cross-section A_0 of the test piece. Any reduction of this cross-section is ignored. The test subjects a sample of material of circular or rectangular cross-section, of a specific gauge length and equipped with end pieces of larger section which taper smoothly to the gauge length.

When subjected to uniaxial tension beyond the limit of proportionality the material within the gauge length elongates plastically, contracts uniformly or locally transversely and work hardens. The stress σ in the material increases but, because of the decrease in the cross-sectional area A, the stress S calculated from the load and the original cross-sectional area A_0 increases more slowly, attains a maximum value S_u and (usually) declines before the specimen breaks.

Limit of proportionality The stress at which elastic behaviour of a material is replaced by a combination of elastic and plastic behaviour, normally expressed as either the *yield stress*, S_y (units MPa), or the *proof stress*, $S_{0.5\%}$, $S_{0.2\%}$, $S_{0.1\%}$ (units MPa), where the departure from elastic behaviour is indicated by the suffix and S (or, in some codes P) is the load:

$$S = \frac{P\left(\text{Yield or strain offset}_{\substack{0.005\\0.002\\0.001}}\right)}{A_0}$$

Ultimate tensile stress S_u (units MPa) The maximum load at which a ductile material fractures in the tensile test divided by the original cross-sectional area. S_u is not to be confused with σ_u which is the true stress:

$$\sigma_u = S_u A_0 A_u^{-1} \tag{1.8}$$

where A_u is the cross-sectional area at the time of failure.

S_u depends on the dimensions of the specimen (the gauge length is normally $0.565\sqrt{A_0}$ but it may be 50 mm or some other value) and the rate of application of stress. Both these parameters should be recorded.

Fatigue endurance Related to S_u rather than S_y. The difference between S_u and S_y is a measure of the safety margin against accidental overload.

Most modern design codes base the permissible stress in a material on a factor (say 66%) of S_y. Some other codes use a factor of S_u as a design criterion. This is cost effective and safe when using a ductile material such as mild steel.

Tensile ductility Reported either as *elongation*, e (units %)

$$e = \frac{\delta}{L} = \frac{L - L_0}{L_0} \times 100 \tag{1.9}$$

where δ is the extension to fracture, or as *reduction in area, A_R* (units %)

$$A_R = \frac{A_0 - A_u}{A_0} \times 100 \tag{1.10}$$

Ductility is the property that confers tolerance to flaws to a material and is also an indication of material quality and correct heat treatment. Standards usually specify a minimum ductility.

Standard procedures for tensile testing are given in BS 18, ASTM E8, ASTM E345 and ASTM B557.

Flexural strength, S (units MPa) The calculated maximum stress on the tensile side of a beam which fails when stressed in bending. It is used to measure the strengths of materials such as cast iron and ceramics which are too brittle to be tested using the standard tensile test. A beam stressed in three-point loading has the maximum stress applied only on one line on the surface. Multiple testing is required to produce results which can be used in design and much higher safety factors (see Section 1.6.10) are required than are used for ductile materials tested using the standard tensile test.

A standard testing method is described in ASTM C580.

1.1.3 Hardness

Hardness is the resistance of a material to permanent deformation by indentation or scratching. It is not a simple intrinsic property of a material but a complex response to a test. Vickers, Brinell and Knoop hardnesses compare the load and the area of the impression produced by an indenter, Rockwell hardness compares the load and the depth of the impression, Shore hardness is a measure of the rebound of an indenter, and Moh hardness measures the ability of one material to scratch another.

Vickers hardness, HV (the dimensions are strictly those of force per unit area, but in practice Vickers and Brinell

hardnesses are comparative numbers), is the quotient obtained by dividing the load F (kgf) by the sloping area of the indentation left in the surface of the material (in mm^2) by a 136° pyramidal diamond indenter:

$$\mathrm{HV} = \frac{2F \sin(136/2)}{d^2} = 1.854Fd^{-2} \quad (1.11)$$

where d is the diagonal of the indentation.

Hardness is a measure of the wear resistance of a material. Used on metals the Brinell hardness value of a medium carbon steel is directly related to the ultimate tensile stress, whilst the Vickers hardness is related to the proof stress. Vickers, Brinell and Rockwell hardnesses can be used to ensure that heat treatment has been carried out correctly.

Hardness testing of ceramics is carried out with very light loads to avoid failure of the material.

Standards for hardness testing are:

Vickers	BS 427, ASTM E92
Brinell	BS 240, ASTM E10
Rockwell	BS 891, ASTM E18
Schlerscope	ASTM 4448

1.1.4 Fracture toughness and impact testing

1.1.4.1 Fracture toughness testing

Plane strain fracture toughness, K_{Ic} (units N-m$^{-3/2}$) The limiting stress intensity required to cause crack extension in plane strain at the tip of a crack when the stress is transverse to the crack. K_{2c} and K_{3c} are parameters corresponding to stresses in the plane of the crack.

Standard testing methods are given in BS 5447 and ASTM E399.

Fracture toughness is sometimes denoted by K_{1c} or K_{Ic}.

Elastic-plastic fracture toughness, J_{Ic} The limiting value of the J integral (which is a line or surface integral used to characterise the fracture toughness of a material having appreciable plasticity before fracture) required to initiate a crack in tension from a pre-existing crack.

Stress intensity to initiate stress corrosion, K_{Iscc} (units N-m$^{-3/2}$) The limiting stress intensity required to initiate propagation of a crack in a specific environment at a specific temperature.

1.1.4.2 Impact testing

In contradiction to fracture toughness testing which quantifies a material property, Izod cantilever and Charpy beam type impact test results are a function of the method of testing. In particular, a machined rather than a fatigue-propagated notch is used. Results are expressed as the energy, J (in joules), required to break the cross-sectional area behind the notch.

Testing a number of specimens of body-centred metals, ceramics and polymers over a range of temperature will reveal a transition temperature below which brittle behaviour is observed. This is reported either as the fracture appearance transition temperature (f.a.t.t. in °C) at which half of the fracture surface is fibrous and half is crystalline, or as the fracture energy transition temperature (in °C) at the inflection in the energy curve. This is a criterion of use for assessing material composition, treatment and behaviour.

Standards for impact testing are BS 131, ASTM E23, ASTM E812 and ASTM E602 (sharp notch tension testing).

1.1.5 Fatigue

S–N curve The graphical relationship between the stress S and the number of cycles N required to cause failure of a material in a fatigue test. This depends on the mean stress, frequency and shape of the stress cycle, the temperature and the environment, all of which should be specified. Note this applies to high cycle fatigue.

High strain fatigue is strain, not stress, related and the plastic strain per cycle resulting in failure ε_p is inversely proportional to $N^{1/2}$ for almost all engineering materials.

Fatigue endurance limit, σ_e (units MPa) The maximum stress below which a material is presumed to be able to endure an infinite number of cycles. This applies only to certain specific engineering materials such as, for example, steel and titanium.

Fatigue limit, σ_{10} (units MPa) The maximum stress below which a material is presumed to be able to endure a specific number of cycles; this is usually of the order of 10^7 to 10^8, but may be lower for specific applications.

The fatigue endurance limit and the fatigue limit are both statistical quantities and depend on the same parameters as the *S–N* curve (see above).

Standard methods for fatigue testing are BS 3518, ASTM E513, ASTM E912, ASTM E206, ASTM E742, ASTM E466, ASTM E606, ASTM 4 468 and ASTM E739. Other ASTM standards are given in *ASTM Standards* Vol. 03;01.

Fatigue life for p% survival (units MPa) The maximum stress below which not less than p% of tested specimens will survive.

Fatigue notch factor, K_f (dimensionless) Ratio of the fatigue strength of a notched to that of an unnotched specimen.

Fatigue notch sensitivity, g (dimensionless)

$$g = \frac{(K_f - 1)}{(K_t - 1)} \quad (1.12)$$

where K_t is the stress concentration factor.

When g approaches 1 a material is fully sensitive; when g approaches 0 a material is insensitive.

1.1.6 Creep and stress rupture

Creep range The temperature range, usually above half the melting-point temperature (in kelvin), at which the design stress computed from creep or stress rupture is lower than that calculated from yield or 0.2% proof.

Stress to rupture, σ_R (units MPa) The tensile stress at which a material will fail if held at a specific temperature for a specific time, depending on the type of application.

Stress to a certain creep strain, σ_ε (units MPa) The tensile stress at which a material will creep to a specific strain ε (ignoring the initial strain on loading) if held at a specific temperature for a specific time. For a specific material σ_ε and σ_R are related.

Creep rupture elongation (units %) The percentage of the original length by which a creep rupture specimen extends before failure.

Larson–Miller parameter, P A parameter used to extrapolate the results of creep rupture tests carried out at relatively short times to longer times. The rate equation is:

$$P = T(\log t_R + C) \quad (1.13)$$

where T is the absolute temperature and C is an empirically determined constant. Other rate equations have been derived by Sherby–Dorn and Manson–Haferd.

Standards for creep and stress rupture testing are BS 5447 and ASTM E1329 for metals and BS 4618 for plastics.

1.1.7 Thermal properties

Specific heat per unit mass, C_p (units J kg^{-1} K^{-1}) The rate of change of heat content of 1 kg of the material with temperature. Specific heats are often quoted in J g^{-1} K^{-1} or in compilations of thermodynamic data as cal mol^{-1} K^{-1}. They may also be quoted as mean specific heats over a range of temperature, usually 25°C to a specific elevated temperature.

Specific heat per unit volume, C_v (units J m^{-3} K^{-1}) The specific heat of a gas at constant volume C_v does not include the work required to expand the gas and is therefore lower than C_p.

Thermal expansion The *linear thermal expansion* α (units K^{-1}) is the fractional increase in length l per degree rise in temperature at a specific temperature T:

$$\alpha(T) = l^{-1}\frac{dl}{dT} \qquad (1.14)$$

More commonly, a *mean expansion* between two temperatures

$$\bar{\alpha} = l_o^{-1}\frac{\Delta l}{\Delta T} \qquad (1.15)$$

where Δl is the change in length from l_o at temperature T_o when the temperature is changed by $\Delta T = T - T_o$, is quoted. In data compilations T_o is often 25°C.

In anisotropic materials (single crystals or materials having a preferred orientation), the thermal expansion coefficient may differ between each of the three orthogonal directions x_i, x_j and x_k.

Thermal conductivity, λ (W m^{-1} K^{-1}) The heat flow per unit area generated by unit temperature gradient:

$$\lambda = A^{-1}\frac{dQ}{dt}\frac{dl}{dT}$$

where dQ/dt is the rate of heat flow across area A and dT/dl is the temperature gradient. λ is normally a function of temperature and, in anisotropic materials, of direction.

Thermal diffusivity, D (units m^2 s^{-1}) A measure of how fast a heat pulse is transmitted through a solid:

$$D = \frac{\lambda}{\rho C_p}$$

where λ is the thermal conductivity, ρ is the density and C_p is the specific heat.

Thermal diffusivity varies with temperature but can be measured more quickly and accurately than thermal conductivity.

1.1.8 Electrical properties

Volume resistivity, ρ (units Ω-m) The resistance (in ohms) of 1 m^3 of a material:

$$R = \rho l A^{-1} \qquad (1.16)$$

where R is the resistance of a body, l is its length and A is its uniform cross-section.

Dielectric breakdown (no standard symbol: units K V mm^{-1} or K V) Measured according to IEC 672, BS 1598: 1964, ASTM D116 or DIN 40685.

Relative permittivity (dimensionless) The ratio of the charge storage capacity of a material in an electric field which results from realignment of the crystal structure compared with the charge storage capacity of empty space.

Permittivity (units A s V^{-1} m^{-1}) Given by:

$(As)/(Vm)$

where A is the current in amperes, s is the time in seconds, V is the electric potential in volts and m is the length in metres.

Dielectric loss, tan δ (dimensionless) The phase angle introduced by the time taken for polarisation to occur on application of a field. Dielectric loss is frequency and (usually) temperature dependent.

1.1.9 Optical properties

Spectral absorption coefficient, K (units mm^{-1}) The log of the ratio of the incident to the transmitted light intensity through unit thickness:

$$K = (\log I_o - \log I)x^{-1} \qquad (1.17)$$

where I_o is the incident intensity, I is the transmitted intensity and x is the thickness (in mm). K varies according to the wavelength of the incident light.

Refractive index, μ (dimensionless) The ratio of the velocity of light in vacuo to that in the medium:

$$\mu = V_{vacuo}V^{-1}{}_{medium} = \sin i \sin^{-1} r \qquad (1.18)$$

where i and r are the incident and refracted angles of the beam to the surface.

1.2 The principles underlying materials selection

1.2.1 Introduction

The need to select a material may arise from a number of circumstances including the following.

(1) An entirely new component is to be developed to perform functions not previously visualised.
(2) A component is required to perform an increased duty which renders the performance of the material previously used unsatisfactory.
(3) The incidence of failure in a material previously specified is too high, or occurs at too early a stage in the life of the component.
(4) Some material shortcoming not strictly related to operational performance has become apparent. A material which was acceptable initially may become unsatisfactory because:
    (a) it has become so expensive, relatively or absolutely, that the equipment, of which the component is a part, can no longer fulfil an economic function;
    (b) it is no longer available locally or globally (or might become unavailable in the event of an emergency); or
    (c) it is no longer acceptable on grounds of health, safety, aesthetics or public sentiment.

Examples of materials which have been developed in answer to the above-listed circumstances are as follows.

(1) The 'magnox' can for the first-generation gas-cooled power reactor.
(2) Superalloy blades of progressively increasing creep resistance culminating (so far) in the directionally solidified castings now used.
(3) Notch ductile aluminium killed steels to replace the materials which failed by brittle fracture in the 'liberty' ships.
(4) (a) Steel-cored aluminium instead of copper conductors for overhead power lines;
    (b) nickel-based superalloys for military jet aircraft after it was realised that the source of cobalt in cobalt-based superalloys were situated in doubtful African or Iron Curtain countries; and
    (c) ceramic fibres to replace asbestos as a binder for heat insulation because of the hazard of 'asbestosis'.

All these examples of material choice were developed by means of the techniques described later in this section. The materials selected have performed entirely satisfactorily, and in those cases where operational parameters were not the cause of replacement the substitute material has in fact performed better than the original.

1.2.2 Techniques of materials selection

There are at least three different techniques by which the optimum material for use in a specific component may be selected.

(1) The *classical procedure* using functional analysis and property specification.
(2) The *imitative procedure* which consists of finding out what material has been used for a similar component.
(3) The *comparative procedure* which consists of postulating that the component be made from some cheap and well understood engineering material, assessing in what ways such a material's performance would be inadequate and from this arriving progressively at the right material.

The classical procedure is the only one that is universally applicable and its use is essential, even when procedure (2) or (3) is followed, to check the findings of the functional analysis and property specification. By itself, however, the classical procedure is expensive and time consuming and requires a considerable amount of prototype testing to ensure that no critical requirement or essential property has been overlooked.

The imitative and comparative procedures, where applicable, provide invaluable shortcuts, save a vast amount of time and money and will help to ensure that no essential parameter has been overlooked. The materials engineer will be wise to employ all three techniques in parallel wherever practicable.

1.2.3 Preliminary examination of design

Whichever of the above procedures is employed, it is essential to commence with an analysis of the function of a component, a critical examination of the design and to establish the property requirements of the material under consideration.

Design affects the materials-selection procedure at all stages. A component may fulfil its function in more than one way due to different designs which result in different materials-property requirements and hence different optimum materials and different manufacturing routes. For example, a box with a hinged lid may be made from two pieces of thin metal sheet and a pin or from one piece of polypropylene.

The effect of design on a manufacturing process is particularly important when considering a materials change in an existing product, for example from metal to plastic or ceramic.

Design and materials selection constitute an iterative process: design affects the optimum material, which in turn affects the optimum design.

1.2.4 The classical procedure

1.2.4.1 Functional analysis

Functional analysis is a formal way of specifying material properties, starting from the function of a component. This involves:

(1) specification of the functions of a component;
(2) specification of the requirements of a component; and
(3) specification of the requirements of the material properties.

1.2.4.2 Function

The overall function should be specified as broadly as possible to allow the greatest number of options in design. Where there are several functions all must be specified. This latter requirement is essential even when the choice of material has been necessitated by the failure of a material to perform one specific function, because a change in a material to make it capable of fulfilling one function may make it incapable of fulfilling another. For example, using a higher tensile steel to carry an increased load may result in brittle fracture under shock.

1.2.4.3 Component requirements

When the functions have been established the component requirements can be identified. For example, the one-piece box mentioned above must be capable of being opened and closed an indefinite number of times. In specifying component requirements it is important to remember that it must be possible to produce the article in the required form, and that the component must withstand the environment in which it is operating, at least for its design life.

1.2.4.4 Materials property requirements

From the component requirements the materials-property requirements can be established. The material for the one-piece box must have an almost infinite resistance to high strain fatigue in air at room temperature. This is obtainable from a polypropylene component manufactured in a specific way.

The property requirements established by the functional analysis may be quantitative or qualitative. For example, the material for an automobile exhaust *must* be sufficiently strong and rigid to withstand weight and gas pressure forces. Quantitative requirements must be established by analysis of the design and operating conditions. In comparison, the requirement to resist corrosion and oxidation is qualitative. Property requirements may also be classified as essential and desirable. The strength requirement in the material for the exhaust is essential. Environmental resistance is often sacrificed to minimise initial cost (even when, as in this example, a more resistant material may be economically superior over the total life of an automobile).

1.2.4.5 Materials requirement check-list

The next stage is the formulation of a 'materials requirement check-list'. Some properties which will feature in this check-

list are listed in Section 1.1. The reader should not be discouraged by the length and complexity of this list. It will in many cases become evident that whole ranges of properties (and materials) may safely be ignored at first glance. For example, if the component is required to transmit or refract light the choice of material is immediately limited to a glass, mineral or polymer and the design and property specification is thus also restricted. If electrical conductivity is significant, choice is limited to conducting metals, resistive or semiconducting materials or insulators.

1.2.4.6 Important characteristics

The important characteristics requiring consideration for many engineering components are:

(1) mechanical properties—stiffness, strength and ductility;
(2) physical properties—thermal, electrical, magnetic and optical properties;
(3) environmental resistance and wear, including applicability of corrosion protection;
(4) capacity for fabrication; and
(5) cost, which includes material, manufacturing, operating and replacement costs.

1.2.4.7 Mechanical properties

Resistance to manufacturing and in-service loads is a requirement of all products. The material must not buckle or break when the component comes under load. A product must also have an economic life in fatigue or under creep conditions. Where a number of materials meet the minimum strength and stiffness requirements, a preliminary short-list can be made on the basis of cost per unit strength or unit stiffness (or in the case of transport applications, strength per unit weight).

1.2.4.8 Physical properties

Physical properties such as, for example, specific gravity, are important for most applications. As noted above, for some applications optical or electrical properties may be paramount.

1.2.4.9 Environmental resistance (corrosion)

Resistance is a property whose universal importance has been obscured by the circumstance that it has been inherent in the choice of materials for most common applications.

Corrosion-resistance requirements vary from the absolute, where even a trace of contamination in a fine chemical, food or cosmetic is unacceptable, to the barely adequate where the cheapest material whose integrity will survive the minimum economic life should be chosen.

When assessing corrosion resistance attention should be paid not only to the rate of general corrosion but also to the possibility of localised corrosion which, as described in Section 1.8, may destroy component integrity without significant dimensional changes.

Corrosion mechanisms may cause the disintegration or deterioration of metals, polymers, ceramics, glasses and minerals.

1.2.4.10 Wear resistance

Wear is the product of the relative movement between one component and another or its environment. Prevention of wear depends principally on design and operation, but can be minimised or eliminated by the correct choice of material, material pair, or coatings.

1.2.4.11 Manufacture and cost

Manufacturing routes are selected on the basis of lowest total cost to produce the desired performance. In the past, performance requirements have favoured certain processes such as, for example, forging instead of casting, but more recently attention to quality improvement techniques in casting has levelled up in-service properties and cost is emerging as the deciding factor.

It is difficult to assess the relative total costs of different material/manufacturing route combinations at the early stage of a design and, wherever possible, finalising precise geometries should be delayed until possible materials and manufacturing routes have been identified, otherwise there will be an avoidable cost penalty.

The cost of a component includes:

(1) in-position costs, which comprise material cost (influenced by quality and quantity), manufacturing cost, quality-control cost and administration cost; and
(2) lifetime costs, which comprise servicing, maintenance, warranty, outage and replacement costs.

Costs which accrue at different periods must be discounted to a common date. Differences in discounting rates between different countries or organisations can lead to the selection of different materials for applications which are, in all other respects, identical.

Comprehensive knowledge of the application is important in assessing the relative importance of cost and performance. Cost is paramount in the case of a widely marketed consumer item where small differences in reliability and life have little influence on saleability, but performance is paramount for certain sporting or military applications. There may, for example, be no advantage to be gained by incurring additional expense to prolong the life of a car exhaust system from 5 to 7 years when the purchaser intends to replace the car after 2 years. In cases such as this the sales department must always be consulted before the final material choice is made. On the other hand the material from which a racing car spring is made must have the maximum possible specific rigidity, regardless of cost. The potential rewards for employing the optimum material far outweight any cost saving which might be obtained by choosing the second-best material.

1.2.4.12 Material selection

When the properties of candidate materials have been ascertained (by the procedures discussed later) a short-list should be established. If it is immediately obvious that one material is outstandingly superior the choice is straightforward. Often there is one property requirement that outweighs all the others. When this is the case the choice is simplified. There may, however, be a number of possible materials, or none may meet all requirements.

A number of procedures have been proposed for eliminating all but one of a number of possible materials. These include an advantage/limitation table, an elimination grid,[1] and ranking methods for properties (and the number of properties) that meet the requirements.

Local factors—using a material which is familiar locally, using a material which has a margin in one specific property that may be of value in a future marque of component, or using a material that is suitable for a locally available fabricating or machining technique—will often influence the final choice.

When no material meets the necessary requirements, a careful re-examination may reveal that a change in design, environment or operating conditions will enable satisfactory

performance at minimum extra cost. As a last resort it may be possible to arrange for easy replacement after a fixed time, and to hold a supply of spares.

1.2.5 Drawbacks of the classical procedure

Application of the procedures outlined above guarantees success if followed logically and completely and if design, operating and material parameters are thoroughly understood. This is seldom the case in practice. Designs cannot always be evaluated precisely, material properties are seldom specified fully, and it is impossible to predict exactly what an operator will do.

In most cases the classical method requires a considerable amount of mechanical-property evaluation, possibly materials and process development and a substantial programme of prototype testing, before satisfactory performance can be guaranteed. Time may not be available to undertake this. The easiest way to reduce the time required is to use the imitative procedure.

1.2.6 The imitative procedure

The imitative procedure involves finding out what material has been used for the same component (or a component as similar as possible) and using this, an improved material, or a material modified for the difference in conditions. Successful implementation of this procedure not only verifies design and reduces the time for materials-property evaluation but also very substantially reduces prototype testing because the most likely causes of failure have already been experienced and cured. The problem is to ensure that the information obtained is accurate, comprehensive and fully understood.

Even within an organisation, operators' reports are not completely reliable. A report of satisfactory performance may merely mean that the operator knows when the component is about to fail so that he can replace it without extra outage. The operator may have found out how to handle this component and a similar component may fail disastrously in the hands of another operator. These difficulties are compounded when information is obtained from an outside source, whether rival or friendly. Informants do not mean to mislead. The information they withhold is usually information that they cannot imagine the recipient does not already possess.

The ability to obtain information when it is required depends on appropriate organisation. Ideally, there should be a materials engineer who combines knowledge of all the materials and requirements of the organisation with an acquaintanceship or, ideally, friendship with all similar persons throughout the world. The right person is, when presented with a problem, able to contact the person who already has experience of the matter wherever she or he may be and obtain the benefit of that experience. His or her knowledge of the other organisation is comprehensive enough to enable him or her to assess the effect of different procedures between the two organisations.

The chemical industry (as described by Dr Edeleanu[2]) operates a world-wide information system with personnel of this type and has found that information on what can be done and how to do it may most efficiently and quickly be obtained in this way.

1.2.7 The comparative procedure

The *comparative* procedure for materials selection involves selecting a cheap, tolerant and well-understood material and investigating to what extent its properties fall short of what is required for the component to operate satisfactorily.

A typical example, and one for which this procedure is extremely suitable, is the specification of a material for chemical process plant.[3] A scheme design is produced using carbon steel which is cheap, readily produced, easily fabricated, ductile and, therefore, tolerant of flaws and geometrical irregularities and corrodes uniformly at a predictable rate. If carbon steel is shown not to be satisfactory the unsatisfactory property or properties can be modified. The necessary change may impair other properties but will do so in a predictable way. Thus:

(1) Improved corrosion resistance may be obtained by the use of a steel with a higher chromium, and possibly a higher nickel, content. This will increase cost and probably also delivery time, render design and fabrication more sensitive and may enhance sensitivity to localised corrosion.
(2) Improved strength may be obtained by the use of a steel with increased carbon and alloy content with drawbacks similar to those that apply in the case of the improved-corrosion-resistance material.
(3) A higher temperature of operation may require the use of a creep resisting steel, again with similar drawbacks to (1).
(4) Operating at a lower temperature may require a steel with guaranteed low-temperature properties or may, at the limit, require an aluminium alloy.

Evidently this procedure, with the exception of the case where a change is made to a completely different material, involves changes which are progressive, and whose effects can be foreseen. Therefore the chances of encountering some unexpected drawback are minimised and the requirement for component testing is minimised also.

1.2.8 Information sources

It has so far been assumed that staff charged with material selection have at their disposal a complete range of information on material properties. This may be the case when electronic databases[4] now being developed are perfected. In the meantime staff should have available for reference British, American and possibly German materials standards, and volumes such as the *ASM Metals Handbook*, the *Plastic Encyclopaedia* and as up to date a ceramic work as is available. In addition, the *Fulmer Materials Optimizer*, which shows properties of all types of engineering materials in the form of comparative diagrams, will prove an invaluable guide to materials selection.

When the field has been narrowed down to a few materials the material manufacturer should be consulted. Organisations such as steelmakers or polymer manufacturers possess more information on their products than has been published and also have experience in their application. They can provide valuable guidance on final selection, design and manufacture. Furthermore, it should be remembered that a standard steel obtained from one manufacturer may differ in some relevant characteristic from the same steel purchased from a competitor. A reputable manufacturer is aware of this and should, in addition to extolling advantages, warn of problems which will have to be overcome.

1.2.9 Computerisation of materials selection

Much effort is at present being deployed towards producing databases of material properties. Three recent international conferences have been devoted to this subject[4] and a directory of databases for materials is available.[5] These databases are not necessarily material-selection systems and much interest

has been directed to providing systems which will undertake material selection using the classical functional-analysis and property-specification procedure.

It is not possible, legitimately, to computerise the imitative procedure because no organisation can be expected knowingly to provide another, possibly competing, organisation with access to programmes intimately concerned with its own design philosophy and development programme.

There is, however, no difficulty in computerising the comparative procedure of materials selection. A computer program which will select the optimum material for a specific application can easily be produced if (a) the materials involved form a very closely related family with very similar properties, (b) no novel and unforeseen failure mechanism takes over, and (c) the properties of the candidate materials have been determined comprehensively. Two such programs are known to exist: ICI (EPOS), for the selection of polymers; and a Sandwik program for selection of cutting tools. These are knowledge-based systems dealing with families of essentially similar materials.

A computer program for selecting process-plant materials (as described in Section 1.2.7) would be equally straightforward, provided the requirements could be met by a steel and no unforeseen failure mechanism took over.

The requirements for a computer program to undertake selection by the classical procedure are much more general and much less well defined. The starting point is a product design specification (PDS) which is a functional and formal statement of what is required from the product to be designed, not a description of the product. The PDS contains a material design specficiation (MDS) which, like the PDS, is incomplete and ill defined. The computer must match this MDS to descriptions of existing materials and materials specifications (MS) which may be incomplete and reflect various levels of confidence. The result of the analysis may be a requirement to modify the PDS, to develop a new material, or to acquire additional information concerning specific materials.

A computer system capable of selecting materials requires:

(1) the ability to deal with simple and complex data structures;
(2) powerful structures for data acquisition and updating by augmentation and modification;
(3) the ability to manage sparse data;
(4) the ability to compare incomplete descriptions;
(5) the ability to distinguish the relationships, and sometimes lack of relationships, between materials, or parameters nominally in the same classification; and
(6) the ability to be easily extensible.

The system must take into consideration:

(1) the duty or function of the component;
(2) the materials properties;
(3) the manufacturing route;
(4) shape, dimensions and failure mode; and
(5) the relative cost of the materials, manufacturing routes and designs considered.

In addition, the system requires certain user characteristics so that it can be operated by designers and engineers and free the materials engineer for long-term difficult and strategic problems. It should:

(1) be rapid in use;
(2) require a minimum of learning;
(3) be accessible at different levels to suit different levels of user;
(4) have text and graphical output; and
(5) have recording facilities.

Various procedures for optimised decision-making have been proposed, including linear programming methods[6] and numerical algorithms.[7] There is a tendency to rely on ranking methods which allocate a rank of 0 to 3 for each material property. This introduces an imprecision which should not be necessary in the application of a computer which should be capable of relating property variation with overall cost.

At least one system is available which is claimed to be applicable at the innovation stage of design. A brief description of this system is given below, as an example of methods which could be employed.

PERITUS[8] is a knowledge-based system which comprises three main stages.

(1) A *director* stage which directs the non-specialist to data and knowledge modules. The structure of this is shown in *Figure 1.1*.
(2) A *presort* stage which uses ranking lists to produce a short-list of candidates from the materials indicated by the director stage (see *Figure 1.2*).
(3) An *evaluation and optimisation* stage which can either display the short-list together with deviations from ideality or, where the required modules exist, uses failure

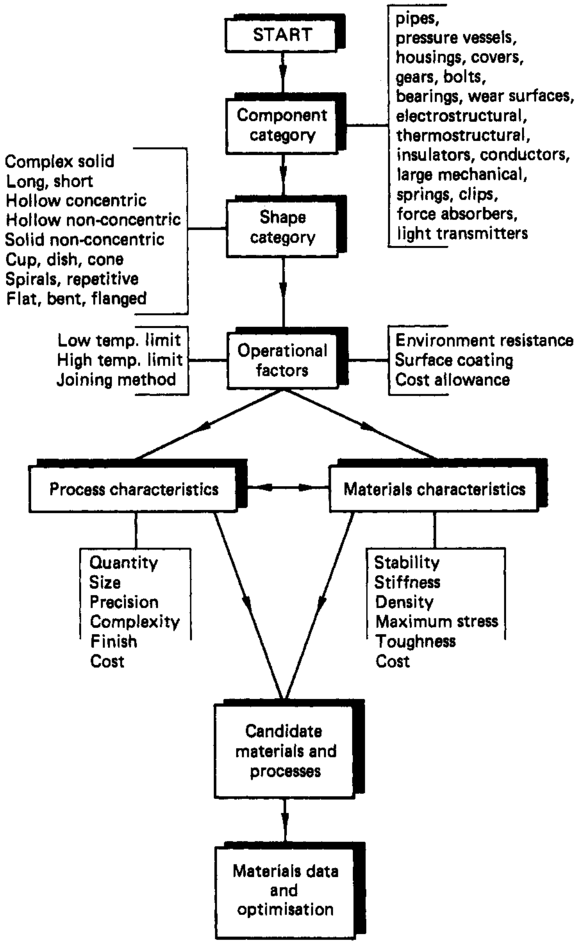


Figure 1.1 The structure and features of the director stage of the PERITUS knowledge-based system for the selection of engineering materials. (Reproduced by permission of *Metals and Materials*)

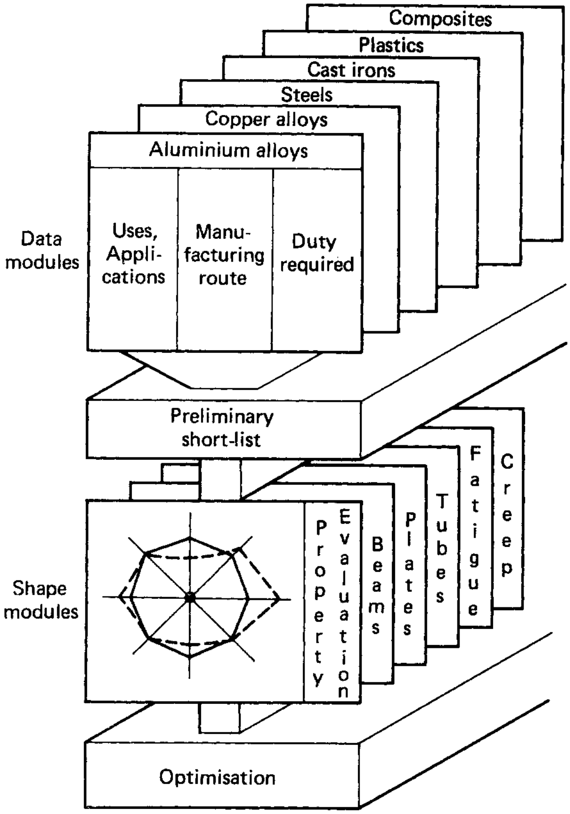


Figure 1.2 A representation of the organisation of the presort, evaluation and optimisation stages of the PERITUS knowledge-based system. (Reproduced by permission of *Metals and Materials*)

modes to optimise component dimensions and properties, presumably including overall cost (see *Figure 1.2*).

This system is modular and additional modules may be added as required or when available.

Further information on the PERITUS system is given by Swindells and Swindells[9] and a conceptual model for materials selections is discussed by Dimaid and Zucker.[10]

1.3 Ferrous metals

1.3.1 Introduction and standards for wrought steels

The development of alloys of iron, which include wrought iron, steel and cast iron, has been essential to a technological society. Iron ore is widespread, cheap and easily mined and from it can be produced alloys with the widest range of properties of any material. Steels may be produced with properties which vary from soft and ductile to strengths and hardnesses which, until the very recent development of the sialons, exceeded those of any relatively inexpensive material.

There are a very large number of wrought steels and numbering systems, German specifications[11] alone include 1400 grades. Because of the rapid evolution of ferrous metallurgy none of these classifications or numbers are ever fully up to date. In engineering, standard steels are commonly referred to by their AISI/SAE number[12] based on composition (*Table 1.1*). The last two figures indicate the carbon content in the case of non-stainless steels. The corresponding general British specification is BS 970:1983 and earlier. In it steels are identified by a three-digit number denoting the type of steel, followed (in the case of non-stainless steels) by a letter denoting the type of specification (A for analysis, H for hardenability or M for mechanical properties) and two numbers denoting carbon content. For stainless steels the first three figures are identical to the AISI figures, the letter is S, and the final two numbers are coded. In the following accounts of classes of steel the AISI/SAE and BS ranges will, where possible, be given. There are also a number of specifications depending on product or application which do not always follow the BS 970 or AISI/SAE numbering system. Apart from the standard steels there are well-known steels which are recognised by designations used by their orginators and are in no way inferior to standard steels. Many steels of this kind are recognised and used world-wide in aircraft specifications and are made by many different steelmakers, although they have not yet been recognised by the several national bodies that govern steel specifications. Cast irons and cast steels have separate specification numbering systems.

1.3.2 Types of wrought steel

There are at least 11 separate classes of wrought steel, most of which are further subdivided.

1.3.2.1 Carbon steel (AISI 1006-1572; BS 970 000-119)

This is the basic type of steel which far exceeds all other metals in tonnage produced. *Low carbon steels* are subdividied into hot-rolled (see Section 1.4.3.1) and cold-rolled steel.

Hot-rolled steel has low strength (although the higher carbon versions can be heat treated to high hardness in small sections) and low toughness, but is readily available at low cost and is easily formed, welded and machined. *Cold-rolled steel* is harder, has good surface finish and dimensional tolerances, high strength at low cost and good machineability.

1.3.2.2 High strength low alloy steels

These are proprietary steels with low carbon made to SAE 950. They have significantly higher strength and are easily formed and welded.

1.3.2.3 Hardened and tempered steels (AISI/SAE 31–98; BS 970 500–599 including higher carbon steels)

These steels contain sufficient carbon and alloy to enable them to be heat treated to the desired strength and toughness at the design thickness. They may have high toughness and high strength at elevated temperature but they are more expensive than carbon steels and the higher alloy steels have poor weldability and machineability.

1.3.2.4 Case hardening steels

These are steels of relatively low carbon content (final BS specification Nos 12 to 25). They may be surface hardened by carburising, carbonitriding or nitriding when heat treatment will produce a very hard surface and a softer (but, where necessary, strong) ductile core. They are used when wear resistance must be combined with core toughness.

2 Polymers, Plastics and Rubbers

John Brydson

Contents

2.1 Introduction 2/3
2.1.1 Polymers, rubbers and plastics 2/3
2.1.2 Thermoplastics and thermosetting plastics 2/3
2.1.3 Amorphous thermoplastics and orientation 2/3
2.1.4 Crystalline thermoplastics 2/4
2.1.5 The glass transition and the crystalline melting point 2/4
2.1.6 The elastic state and vulcanised rubber 2/4
2.1.7 Thermoplastic rubber 2/5
2.1.8 Additives 2/5
2.1.9 Polymer blends 2/6
2.1.10 Composites 2/6

2.2 General properties of rubbery materials 2/6
2.2.1 Stress–strain properties in tension 2/6
2.2.2 Modulus and hardness 2/7
2.2.3 Bulk modulus, Poisson's ratio and incompressibility 2/7
2.2.4 Stress–strain properties of rubber in shear 2/7
2.2.5 Stress–strain properties of rubber in compression 2/7
2.2.6 Resilience 2/9
2.2.7 Creep, set and dynamic properties 2/10
2.2.8 Non-rubbery properties of technical importance 2/10

2.3 Survey of commercial rubbery materials 2/11
2.3.1 Tyre rubbers 2/13
2.3.2 Special purpose rubbers 2/14
2.3.3 Speciality rubbers 2/15
2.3.4 Thermoplastic rubbers 2/15
2.3.5 Liquid processing rubbers 2/15
2.3.6 Exotics 2/15

2.4 General properties of plastics 2/15
2.4.1 Softening point 2/16
2.4.2 Thermal stability 2/16
2.4.3 Flame retardancy 2/17
2.4.4 Toughness 2/18
2.4.5 Creep behaviour 2/19
2.4.6 Solvent resistance 2/20

2.5 Survey of commercial plastics materials 2/22
2.5.1 Major tonnage thermoplastics 2/22
2.5.2 Engineering thermoplastics 2/24
2.5.3 Other thermoplastics 2/25
2.5.4 Thermosetting plastics 2/25

2.6 Processing of rubbers and plastics 2/26
2.6.1 Shaping and setting 2/26
2.6.2 Processing in the molten state 2/26
2.6.3 Compression moulding 2/28
2.6.4 Transfer moulding 2/28
2.6.5 Injection moulding 2/29
2.6.6 Extrusion 2/31
2.6.7 Blow moulding 2/33
2.6.8 Calendering and frictioning 2/35
2.6.9 Thermoforming 2/35
2.6.10 Other processing methods 2/36

2.7 Design of rubber components 2/37
2.7.1 Rubber springs 2/38
2.7.2 Rubber seals 2/41
2.7.3 Vibration isolation 2/41

2.8 Design of plastic components 2/43
2.8.1 Designing under creep loading 2/43
2.8.2 Stiffness of structural foam mouldings 2/44
2.8.3 Design of plastic parts 2/44

2.1 Introduction

2.1.1 Polymers, rubbers and plastics

Rubbers and plastics are important examples of materials based on chemicals known as *high polymers*. Other materials which belong to the same family include adhesives, surface coatings and fibres.

Such polymers are characterised by being made up of a large number of repeating units. One such example is polyethylene which is usually made by joining together many molecules of ethylene, a gas consisting of two atoms of carbon and four of hydrogen and which has the structure shown on the left-hand side of the following diagram. (In this diagram 'C' represents a carbon atom and 'H' one of hydrogen, a single link represents a single bond and the double link between the two carbon atoms a double bond.)

```
H  H
|  |
C=C → —CH2—CH2—CH2—CH2—CH2—CH2—
|  |
H  H
```

Ethylene is an example of a *monomer* and the process of joining up the molecules is known as *polymerisation*. In this case chemical reactions are initiated that lead to the double bonds opening up so that the particular process is often known as *double-bond polymerisation*. Many other important plastics such as polyvinyl chloride (PVC), polypropylene and polystyrene are made in a very similar way.

In the above examples only one monomer is used to make the polymer and the products may be referred to as *homopolymers*. Where more than one monomer is used the process is known as *copolymerisation* and the products as *copolymers*. In the case of two monomers the products are strictly known as 'binary copolymers' (although the adjective is commonly dropped), whilst with three monomers the products are known as 'ternary copolymers' (or more commonly simply as *terpolymers*).

Other ways of making polymers also exist. Without going into detail, the polyamides (nylons) are made either by opening up small molecules which have a ring structure (*ring-opening polymerisation*) or by reacting small molecules which become joined together with the elimination of some small molecule. This is an example of *condensation polymerisation*, in this case the small molecule being split out is that of water.

To say that a chemical is a 'high polymer' means that it has a high relative molecular mass (which is also known as 'molecular weight'). This is approximately the number of times the molecule is heavier than an atom of hydrogen. Common laboratory chemicals usually have molecular weights of less than 300. For example, that of water is 18, benzene 78, ethyl alcohol 46 and sulphuric acid 98. Typical commercial polymers may have molecular weights ranging from a few thousand to over a million. A typical grade of polystyrene, for example, has a molecular weight of about 200 000.

Although it is the high molecular weight which confers on polymers important properties it is as well to bear in mind that, since the weight of an atom of hydrogen is only about 1.7×10^{-24} g, then the absolute weight of a polymer molecule is still very small.

A further point that should be made at this stage is that rubbers and plastics do not solely consist of polymers. It is usual to incorporate a number of *additives* into the polymer in order to adjust the properties. For example, PVC plastics may contain plasticisers (to increase flexibility), stabilisers (to improve heat and light resistance), lubricants (to aid flow or to stop sticking to processing equipment), fillers (such as china clay, usually primarily to reduce cost) and colorants.

2.1.2 Thermoplastics and thermosetting plastics

All of the polymers mentioned in the previous section are long chain-like polymers (generally referred to by chemists as *linear polymers*). Let us consider what happens when we heat up and subsequently cool one of these, namely polystyrene. At usual ambient temperature polystyrene is hard and rigid. The long-chain molecules are intertwined and are generally fixed in space, rather like a frozen mass of spaghetti. If the polystyrene is then heated up to above 100°C the molecules acquire energy which enables them in effect to 'wriggle about'. If subjected to an external stress such as a shear stress or simply the effect of gravity, the mass starts to move or, in common terminology, we say that the material has melted. At this stage the material may be shaped by such processes as moulding or extrusion. On cooling down the mass it will harden again. Materials that behave in this way are known as *thermoplastics*. In theory, heating and cooling processes can be repeated over and over again. However, polymers can be damaged to a greater or lesser extent each time they are heated and sheared, so that there is a limit to the number of times the process may be repeated.

Until about 1960, thermoplastics were less important than *thermosetting plastics*. As supplied to the processor, these materials contain fairly small polymer molecules. However, during and after shaping operations, reactions occur which cause the molecules to join up into a three-dimensional network and the polymer becomes hard or *set* at the processing temperature. This process is generally irreversible and is frequently referred to as *cross-linking*. (It may be noted at this point that, although the term 'thermosetting' was originally used to denote materials that were set at elevated temperatures a number of similar materials may now be hardened at normal ambient temperatures. Strictly, it is probably better to use the term *cross-linked plastics* to cover these materials, irrespective of the reaction temperature used.) Important thermosetting plastics include the phenolic resins, urea formaldehyde and melamine formaldehyde resins, and epoxide resins. Some polyesters are thermoplastic and some are thermosetting (and, for that matter, some are rubbers and some are fibres).

2.1.3 Amorphous thermoplastics and orientation

It is important to distinguish between two types of thermoplastic: the amorphous and the crystalline types. In the former type, exemplified by polstyrene and acrylic plastics such as polymethyl methacrylate, the polymer molecules will, given time, coil up into a random structure in the melt state. When the melt is cooled in the absence of external stresses this random structure will be retained. This leads to the following characteristics:

(1) the solid material will be isotropic in its properties (i.e. they will be the same in every direction);
(2) the shrinkage that occurs on solidification will be small (typically about 0.005 cm per centimetre in an injection-moulding process); and
(3) in the absence of powdery additives or bubbles of air or other gas the product is likely to be transparent.

In practice, the molecules are often subject to extensive shear in the melt during processes such as injection moulding and extrusion, and the melt freezes in an oriented or partly

oriented state so that the product is anisotropic. The following types of orientation may be distinguished.

(1) Monoaxial orientation—this occurs when a molten rod is elongated into a filament. The molecules tend to align with the filament axis leading to increased strength and modulus in the axial direction.
(2) Biaxial orientation—this occurs when a sheet or film is stretched in two directions in the plane of the sheet. This can give a toughening effect to the sheet.
(3) Triaxial orientation—because solid thermoplastics are very resistant to volume change triaxial orientation leading to general dilatation is virtually impossible but locally may cause fracture.
(4) Complex orientation—this commonly occurs during injection moulding where orientation tends to follow the lines of melt flow and thus varies from place to place.

Orientation may or may not be a desirable feature as will be discussed in Section 2.8.

2.1.4 Crystalline thermoplastics

Several thermoplastics in the solid state show many of the characteristics of crystalline materials. In such cases individual molecules pass through several regions where segments of many molecules are arranged or packed in a highly ordered way. Such ordered regions tend to be spatially mingled with less ordered and virtually amorphous regions. Examples of crystalline polymers are polyethylene, polypropylene, the nylons and the polyacetals. As a result of this ability to pack, crystalline polymers show the following general characteristics:

(1) there is a greater shrinkage on cooling from the melt (typically 0.015–0.060 cm per centimetre), which depends on polymer type and the processing conditions;
(2) properties can be affected by seeding to encourage initiation of crystallinity;
(3) because the different zones will have different densities and hence different refractive indices crystalline polymers tend to be opaque although there are some exceptions; and
(4) since levels of crystallinity vary with temperature and with processing conditions, the physical properties of products made from crystalline thermoplastics are more dependent on the manner in which they have been processed and on the ambient temperature than are those of amorphous thermoplastics.

As with the amorphous thermoplastics, when cooled from the molten state without subjection to external stress crystalline thermoplastics are isotropic in that the crystalline structure tends to be randomly dispersed. Furthermore, orientation of the various types indicated with amorphous thermoplastics can also occur. In the case of crystalline thermoplastics, however, orientation of the crystal structures is more important than orientation of molecules. Most synthetic fibres are made by drawing crystalline thermoplastic materials. Because of the layering effect that takes place it is interesting to note that biaxial stretching tends to enhance the transparency of crystalline materials.

2.1.5 The glass transition and the crystalline melting point

Successful processing and use of plastics and rubbers are facilitated by an understanding of two properties: the *glass transition temperature* and the *crystalline melting point*.

At room temperature an amorphous thermoplastic material such as polystyrene is hard and rigid because the constituent molecules are also effectively frozen in space. If, however, polystyrene is heated up to above 90°C the material becomes more flexible and further heating takes it into a rubbery state. If the molecular weight is not too high, raising the temperature above about 150°C causes the material to melt but, in the case of very high molecular weight materials (e.g. over 500 000), the material may decompose before melting as the temperature is raised. The change at about 90°C arises because the polymer molecules start to 'wriggle' and on application of a stress they tend to be elongated. However, entanglements and limited motion stop the chains from sliding past each other so that, on release of a stress, the molecules tend to coil up randomly and the sample recovers, at least in part, its original shape. In other words its exhibits *rubberiness* (also known as 'reversible high elasticity'). This transition from being glass like to rubbery is known as the *glass transition temperature*, universally given the abbreviation T_g. Many properties other than mechanical ones change at the glass transition temperature whilst processing behaviour is very dependent on the relationship between processing temperatures and T_g. It may be noted that T_g does depend slightly on molecular weight and also on the method by which it has been measured, so that any quoted figure can only be considered as approximate. However, for practical purposes, this does not matter too much.

In the case of crystalline polymers the changes at T_g are less, the extent depending on the degree of crystallinity. Indeed, with highly crystalline materials T_g is often difficult to identify. This is because many of the polymer molecules are incorporated in whole or in part in crystalline structures which still exist above T_g, thus still effectively freezing the mass. In the case of lightly crystalline structures the crystal structures present act more like cross-links between which the molecules are able to coil and uncoil on application of a stress so that a measure of rubberiness may be observed.

If, however, the crystalline polymer is progressively heated to higher temperatures, it may be observed that the degree of crystallinity will fall. Eventually a temperature is reached where the last traces of crystallinity disappear and this is known as the *crystalline melting point* (T_m). (In practice, it is becoming more common to obtain this value from peaks observed in differential scanning calorimetry which may be slightly lower.)

It is frequently found that the following rule-of-thumb relationship exists between T_g and T_m

$T_g = 2T_m/3$ (when expressed in degrees Kelvin)

In many common instances this means that the crystalline melting point is about 100 K above the T_g. Providing the molecular weight is not so high that entanglements prevent flow above this temperature, the polymer rapidly becomes molten rather than rubbery. The various relationships are illustrated schematically in *Figure 2.1*.

At this point it is also instructive to show graphically how the stiffness or modulus of the various types of material considered so far are affected by temperature. This is shown in *Figure 2.2*.

2.1.6 The elastic state and vulcanised rubber

It has already been indicated that high elasticity in polymers arises from the fact that, on application of a stress, molecules above T_g are able to deform from a randomly coiled up state. Where there is some entanglement, flow or slippage of molecules past one another is difficult so that when the stress

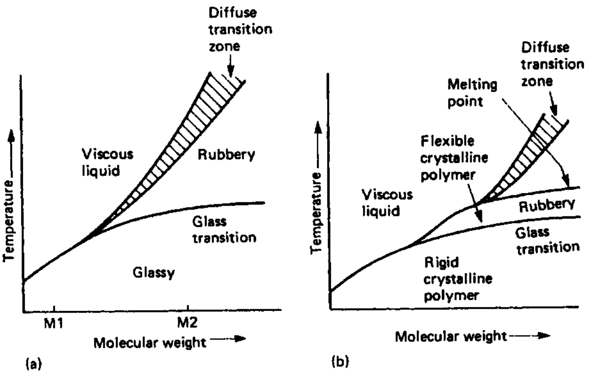


Figure 2.1 Temperature vs. molecular weight diagrams for (a) amorphous and (b) moderately crystalline polymers (with highly crystalline polymers the glass transition is less apparent). (Reproduced by permission from Brydson, *Plastics Materials*, Butterworth)

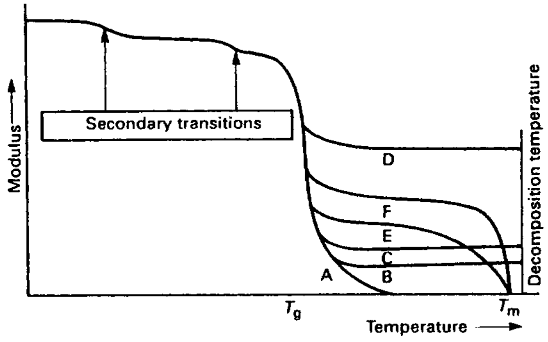


Figure 2.2 Scheme to show the dependence of the modulus of a polymer on a variety of factors. A is an amorphous polymer of moderate molecular weight, whereas B is of such a high molecular weight that entanglements inhibit flow. Similar effects are shown in C and D, where the polymer is lightly and highly linked, respectively. In E and F the polymer is capable of crystallisation, F either being more highly crystalline than E or containing fibre reinforcement. (Reproduced by permission from Brydson, *Plastics Materials*, Butterworth)

is removed the molecules recover a coiled up state. Simply relying on entanglements or small crystalline structures acting as knots is somewhat inefficient and substantial slippage or creep may occur. This may be very much reduced by introducing a small amount of cross-linking, as with thermosetting plastics, but to a much lower extent than prevents chain slippage. In rubber technology this process is known as *vulcanisation*. It is interesting to note that the process was first discovered by Charles Goodyear in the USA as long ago as 1839 when natural rubber was heated with sulphur, a process exploited first by Thomas Hancock in London. Today the use of 1–2 parts of sulphur per 100 parts of rubber is still by far the most common way of vulcanising both natural and most of the large tonnage synthetic rubbers.

In effect the cross-links tend to lock the polymer chains relative to each other in space so that on release of any deforming stress the mass recovers its original shape. Unfilled vulcanised rubbers may be stretched more than 1000%, hence the term 'high elasticity' as distinct from the 'elasticity' of metals. Commercial rubbers, however, usually have much lower breaking elongations (150–550%) in order to obtain improvements in other properties such as stiffness, tear resistance and abrasion resistance.

2.1.7 Thermoplastic rubber

One of the more interesting developments in the past 30 years has been that of thermoplastic rubbers. One of the problems with conventional vulcanised rubber is that, once cross-linking or vulcanisation has taken place, the material is, like a thermosetting plastic, 'set' and cannot be melted and reprocessed. Thus material used in a process but not incorporated into the final product or defective products cannot be re-used as can a thermoplastic material. Thus many attempts have been made to produce a rubbery material in which effective cross-links exist at normal use temperature but which disappear (become *heat fugitive*) at elevated temperatures. To some extent entanglements and crystalline zones fulfil this role and rubbers have been used that are simply very high molecular weight polymers (but thus difficult to process) or slightly crystalline. Polymers have also been made in which ionic cross-links exist at low temperatures, but which lose their force at elevated ones. A number of true covalent-bonded systems have been devised in which these bonds break down at elevated temperature, but which reform at low temperatures. Such systems have been of limited use. Far more successful has been the use of *block copolymers*.

Such block copolymers differ from the more common *random copolymers* in that the monomers of each type are grouped together in one chain. One such material consists of a block of butadiene molecules (forming a segment of a rubbery polybutadiene block) set between two blocks of styrene molecules (forming glassy polystyrene blocks). Such as system is known as a *styrene–butadiene–styrene (SBS) triblock copolymer*. At room temperature the polystyrene ends congregate into domains effectively forming cross-links between many triblock molecules at the chain ends. However, above T_g these domains tend to break up and, because the overall molecular weight is quite low, the whole system melts and is capable of flow. When the melt is cooled the domain structures and thus the cross-links reform. Slightly different are the polyether–polyester block copolymers in which amorphous polyether zones are separated by crystallisable polyester blocks. At room temperature these latter blocks do crystallise together to produce small crystal structures which act as cross-links. These latter materials are available in a variety of polyether/polyester ratios and thus vary in stiffness and rubberiness. Because of the high melting point of the polyester blocks these materials have good heat resistance, and because of their chemical nature they have good oil resistance. They have become important engineering rubbers.

Blends of polypropylene (which is a crystalline polymer) with rubbery ethylene–propylene rubbers, which are very slightly crystalline have also been prepared. Appropriately modified by a heat treatment these blends have also been successful as thermoplastic rubbers.

2.1.8 Additives

The polymers used for making rubbers and plastics are very rarely used on their own without the incorporation of additives. In the case of a rubber it is usually necessary to incorporate a *vulcanising agent* such as sulphur. Since the reaction between most rubbers with sulphur is slow and inefficient, a chemical known as an *accelerator* is often used,

an additive which often works better in the presence of an *accelerator activator* such as zinc oxide and a fatty acid such as stearic acid. To retard ageing, *antioxidants* and *antiozonants* may be used. To increase stiffness, tear resistance and abrasion resistance *reinforcing fillers* such as fine particle carbon blacks may be incorporated, whilst coarser, so-called *inert fillers* may be used primarily to reduce cost. Adjustment of stiffness may be made by softeners (plasticisers), whilst blowing agents may be used to give a cellular structure. Non-black compounds may incorporate colorants. Thus, for example, a bridge bearing based on natural rubber could consist of the following formulation:

Rubber	Natural rubber	100
Vulcanising system	Sulphur	2.5
	Zinc oxide	5
	Stearic acid	2
	Accelerator (e.g. CBS)	0.6
Reinforcing filler	General purpose furnace type (GPF) carbon black	50
Protective system	Antioxidant (TMQ)	3
	Antiozonant (IPPD)	2
	Paraffin wax	2

where CBS is *N*-cyclohexylbenzothiazole-2-sulphenamide, IPPD is *N*-isopropyl-*N*′-phenyl-*p*-phenylene diamine, and TMQ is polymerised 1,2-dihydro-2,2,4-trimethylquinoline.

As one example with thermoplastics, a compound for PVC wire covering insulation flex would contain, in addition to the polymer, *plasticiser(s)* to make the material flexible, china clay as a *filler* which also improves electrical insulation properties, a *stabiliser* to reduce degradation on heating and exposure to light, a *colorant* and a *lubricant* (to prevent sticking to processing machinery). For both rubbers and PVC other additives may also be used for other purposes and similar comments also apply for other rubbers and plastics.

2.1.9 Polymer blends

In recent years increasing use has been made of blends of polymers. In many cases the reason for this is to produce tough polymers and frequently this involves blending a glassy amorphous thermoplastic with a rubbery material as, for example, with toughened polystyrene, a blend of glassy polystyrene with rubbery polybutadiene, and the acrylonitrile–butadiene–styrene (ABS) plastics, originally blends of glassy acrylonitrile–styrene with butadiene–acrylonitrile, but today rather more complex structures. Other important blends include blends of polystyrene and the crystalline polymer 2,6-dimethyl-*p*-phenylene oxide (the produce marketed as Noryl) and blends of polypropylene and ethylene–propylene rubbers to give a type of thermoplastic rubber.

In principle, these materials, often known as *polyblends*, are a physical mixture of different polymers and are thus distinct from copolymers where the different monomers are on the same polymer chain. Therefore a blend might be expected to have quite different properties from a copolymer. For example, a random copolymer usually has too irregular a structure to enable it to crystallise, whereas it is often possible to make a polyblend containing one or more crystalline polymers so that the polyblend retains a measure of crystallinity. In practice this situation may be more complicated with some blends involving copolymers, whilst in some instances there may be chemical reaction between the components of a blend.

2.1.10 Composites

By definition a *composite* consists of two or more physically distinct materials which are combined in a controlled way. Three classes of composite may be distinguished:

(1) where one material (component) forms a continuous phase (*continuous matrix*) and the other components are embedded within this matrix;
(2) *laminar* structures where the components are in separate layers (alternating where there are just two components); and
(3) *interpenetrating* systems where both components form a continuous matrix.

Strictly speaking any polymer compound containing an additive or a polyblend could be considered a composite, although common usage of terminology excludes these materials.

Whilst all three types are used the most typical in the case of rubbers and plastics is the first group and, specifically, the *structural composites* which are characterised by the matrix material being enhanced in stiffness and strength by the presence of a *reinforcing material* embedded within it. Where the continuous phase is a polymer, such materials are known as *polymer-based structural composites*.

Such systems have long been used in the rubber industry where, for example, the strength of a tyre is greatly enhanced by incorporating fibrous filament or cloth which may be organic (e.g. cotton) or a synthetic fibre, or inorganic (e.g. glass fibre or metal filament).

In the plastics industry, laminates using paper have long been used for electrical insulation, whilst cloth-based phenolic laminates have been used for constructional materials for most of the twentieth century. Rather better known is the use of glass-fibre reinforcement in polyester resins for sports cars, boats and railway engine cabs and in epoxide resins in the aircraft and chemical industries. For extreme operating demands the reinforcement may be carbon fibre or an *aramid* fibre (aromatic polyamide fibre).

2.2 General properties of rubbery materials

In this section the general properties of rubbery materials relevant to their use in engineering applications are considered, followed in the next section by a short review of the types available. Unless otherwise stipulated, the use of the word 'rubber' implies vulcanised rubber.

Today, the bulk of manufactured rubber articles is used in engineering applications such as tyres, springs, seals, bearings, mine belting, transmission belting, chemical plant liners and electrical cable and wire insulation in severe environments. Successful use clearly requires an understanding of relevant properties.

2.2.1 Stress–strain properties in tension

Probably the best known property of rubbers is that they are capable of large elastic deformations. It is also well known that this deformation is highly recoverable on release of stress. Two other properties may be noted at this stage:

(1) rubbers have a low modulus or stiffness;
(2) rubbers are virtually incompressible in bulk (Poisson's ratio ν is about 0.5).

In tension an unfilled natural rubber vulcanisate (the product of vulcanisation) may be extended in tension by more than 1000%. Commercial products which more commonly contain

large amounts of filler will have lower *elongations* at break, but for many engineering applications values do exceed 400%. The tension stress–strain curve is not linear but is somewhat S shaped (*Figure 2.3*).

A number of attempts have been made to derive theoretical expressions to fit this curve. The so-called 'Gaussian theory' yields the expression

$$f = \tfrac{1}{2}G(\lambda_1 - 1/\lambda_1^2)$$

where f is the tensile stress, G the shear modulus and λ the strain (strained length/original length).

This gives fair fits up to about 500% elongation, but is poor at higher elongations. A rather better fit is given by the Mooney–Rivlin equation

$$f = 2(\lambda - 1/\lambda^2)(C_1 + C_2/\lambda)$$

where C_1 and C_2 are constants.

In practice, most uses of rubber in engineering applications involve elongations of less than 50%. In this range the rubber is almost Hookean in behaviour and it is possible to use a Young's modulus derived from the slope within this range.

Compared with metals the modulus of rubbery materials is very low. Some typical figures are given in *Table 2.1*.

2.2.2 Modulus and hardness

Rubber technologists make wide use of *hardness measurements*. With rubbers this involves the measurement of a reversible deformation produced by a specified indentor, unlike metal hardness measurements which are concerned with an irreversible plastic indentation. There is a good correlation between hardness and Young's modulus. The most widely used hardness test is that which yields 'international rubber hardness degrees' (IRHD). Also often quoted in the literature are British Standard (BS) hardness degrees and Shore A units. All three give similar numerical values. *Table 2.2* shows the relationship between various types of modulus and IRHD.

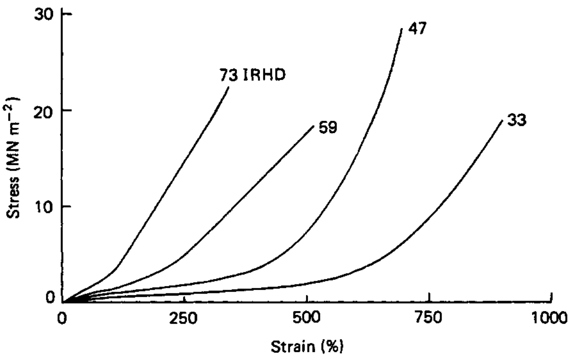


Figure 2.3 Tensile stress – strain curves for four natural rubber compounds of different hardness: 73 IRHD contains 50 parts of a reinforcing black; 59 IRHD contains 50 parts of a non-reinforcing black; and different vulcanising systems account for the different curves of the two gum compounds (47 and 33 IRHD). (Reproduced by permission from Lindley, *Engineering Design with Natural Rubber*, Malaysian Rubber Producers' Research Association)

Table 2.1 Comparison of the modulus and Poisson's ratio of rubbers with those of mild steel

	Natural rubber		
	*Gum**	*Filled†*	*Mild steel*
Young's modulus (MPa)	1.8	5.9	210 000
Shear modulus (MPa)	0.54	1.37	81 000
Poisson's ratio	0.4997	0.4997	0.29

* A vulcanisate of an unfilled natural rubber compound.
† A vulcanisate.

2.2.3 Bulk modulus, Poisson's ratio and incompressibility

The bulk modulus of rubber is in the range 1000–2000 MPa and Poisson's ratio is almost exactly 0.5. This means that solid rubber vulcanisates are virtually incompressible (in bulk). *It is therefore essential that if rubber is being used in an application which involves deformation, there should be some room for the rubber to deform.* For example, if a rubber disc was placed at the base of the inside of a cylinder and a close-fitting piston was dropped into the cylinder, the piece of rubber would be unable to deform and would behave as a hard solid. Failure to realise this fact of incompressibility even with very soft rubbery articles has led to embarrassing design failures. (This restriction does not, of course, apply to cellular (sponge) rubber articles.)

2.2.4 Stress–strain properties of rubber in shear

At constant rates of shear deformation the stress–strain relationship in shear is linear up to about 100% shear strain, i.e. Hookean behaviour may be assumed in shear. In this case it is possible to postulate the relationship

$$\tau = G \tan \gamma$$

where τ is the shear stress and $\tan \gamma = x/t$ where x is the displacement in the direction of shear and t is the separation between shearing surfaces (*Figure 2.4*).

It is also possible to define the shear stiffness K of a mounting being sheared by

$$K = F/x = GA/t$$

where F is the shearing force and A the area over which the force is applied.

It may be noted at this stage that inserting plates into a rubber mounting parallel to the shear direction does not affect the stiffness of the plate, providing that the total thickness of rubber in the mounting is unchanged. The importance of this will become apparent later.

2.2.5 Stress–strain properties of rubber in compression

The compression stress–strain properties of vulcanised rubber are usually measured by compressing a block of rubber between two metal plates. Unless special measures are taken to lubricate the interfaces between rubber and metal, the rubber at the interface will be constrained. Thus, on compression the interfacial area will not change but, since rubber is incompressible in bulk, there must be a lateral expansion between the plates which is at a maximum at the midplane of the sample. In turn this will mean that the free surface of the rubber block will increase, setting up tensile strains on the surface (*Figure 2.5*).

This tensile strain will be in addition to the forces required simply to compress the sample and will, therefore, increase the stiffness. It may be appreciated that this effect is greater

Table 2.2 Hardness and elastic moduli*†

Hardness‡ (IRHD ± 2)	*Young's modulus§*, E_0 (MN m^{-2})	*Shear modulus//*, G (MN m^{-2})	k**	*Bulk modulus*, E_∞ (MN m^{-2})
30	0.92	0.30	0.93	1000
35	1.18	0.37	0.89	1000
40	1.50	0.45	0.85	1000
45	1.80	0.54	0.80	1000
50	2.20	0.64	0.73	1030
55	3.25	0.81	0.64	1090
60	4.45	1.06	0.57	1150
65	5.85	1.37	0.54	1210
70	7.35	1.73	0.53	1270
75	9.40	2.22	0.52	1330
Shore A (approx.)	(lbf in.$^{-2}$)	(lbf in.$^{-2}$)		(lbf in.$^{-2}$)
35	168	53	0.89	142 000
45	256	76	0.80	142 000
55	460	115	0.64	154 000
65	830	195	0.54	171 000
75	1340	317	0.52	189 000

* Reproduced by permission from *Engineering Design with Natural Rubber*, 4th edition, Malaysian Rubber Producers' Research Association.
† Based on experiments on natural rubber spring vulcanisates containing (above 48 IRHD) SRF black as filler. Note that hardness is subject to an uncertainty of about ±2 IRHD.
‡ The majority of springs are in the hardness range 40–60 IRHD.
§ Theoretically, with a Poisson's ratio of ½, E_0 should equal $3G$. This is so for soft gum rubbers, but for harder rubbers containing a fair proportion of non-rubber constitutents, thixotropic and other effects increase E_0 to about $4G$.
// Average design limits: 15% compression, 50% shear.
** k is used in the calculation of compression characteristics.

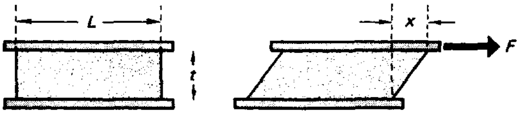


Figure 2.4 Rubber mounting subjected to shear deformation

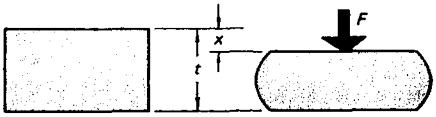


Figure 2.5 Rubber block in compression.

with thin blocks and it is necessary to introduce a *shape factor* S in order to determine the compression modulus E_c effective in any particular sample, test piece or mounting. This shape factor is given by

$$S = LB/2t(L + B)$$

for a rectangular block of length L, breadth B and thickness t.

For a block of square section (i.e. $L = B$) or circular section (diameter L) then

$$S = L/4t$$

The compression modulus depends on the shape factor by the relationship

$$E_c = E_Y(1 + 2kS^2)$$

where E_Y is Young's modulus and k is a numerical factor the values of which are given in *Table 2.2*.

The compression stiffness K_c may be defined as

$$K_c = F/x = E_cA/t$$

where x is the deflection.

The load (F) vs. deflection (x) curve for vulcanised rubber is non-linear. Providing there is no slip at the interface it approximates to

$$F = E_cAe(1 + e)$$

where e is the compressive strain ($= x/t$). The non-linearity is usually ignored for strains up to 10% (*Figure 2.6*).

If plates parallel to the metal–rubber interfaces are inserted into the blocks, one thick sample will in effect be replaced by a number of thinner samples with a higher shape factor (*Figure 2.7*). In effect we have a stacked array of rubber springs. If these springs have deflections x_1, x_2, x_3, etc., the total deflection is

$$x_1 + x_2 + x_3 + \ldots$$

Since the deflecting force will be the same for each element, then the total stiffness for the unit K_{tot} is given by

$$1/K_{tot} = 1/K_1 + 1/K_2 + 1/K_3 + \ldots$$

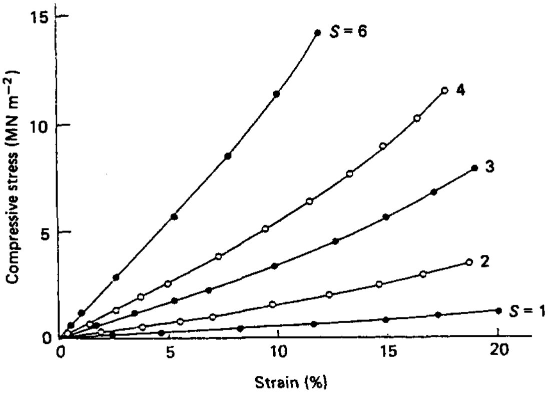


Figure 2.6 Effect of shape factor: experimental stress–strain curves for 6.3 mm thick discs of rubber (47 IRHD) in compression. The shape factor is shown alongside each curve; the diameter (in millimeters) is 25.4 times the shape factor. (Reproduced with permission from Lindley, *Engineering Design with Natural Rubber*, Malaysian Rubber Producers' Research Association)

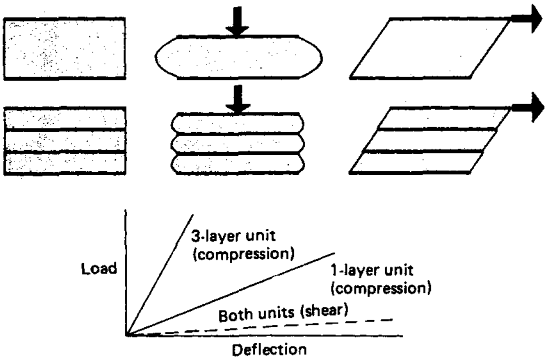


Figure 2.7 The vertical stiffness of a rubber block can be increased by inserting into the block horizontal metal spacer plates which reduce the freedom of the rubber to bulge. The shear stiffness is not altered by the presence of these horizontal plates. (Reproduced by permission from Lindley, *Engineering Design with Natural Rubber*, Malaysian Rubber Producers' Research Association)

If all the rubber 'springs' between the plates are equal and there are n separating metal plates, i.e. $(n + 1)$ rubber springs, then the expression simplifies to

$$K_{tot} = K/(n + 1)$$

where K is the compression stiffness of each element.

Simple calculations show that insertion of such metal plates markedly increases the compression stiffness. For example, *Table 2.3* illustrates how the insertion of a number of metal plates increases the stiffness of a rectangular mounting of loaded face area L^2 and thickness $t = L/4$. This has the effect of *increasing the compression stiffness but not affecting shear stiffness*. This is of considerable importance as a means of producing rubber components with differing levels of stiffness in different directions. This is discussed in Section 2.7.1.3.

Table 2.3 Effect of the insertion of metal plates on the compression stiffness of a rubber mounting of loaded face area L^2 and thickness $L/4$

No. of plates	*S*	*Increase in stiffness,* K/K_0*
1	1	3
2	2	9
3	3	19
5	5	51
7	7	99
10	10	201

* K is the stiffness of the composite and K_0 is the stiffness of the block with no inserted metal plates.

2.2.6 Resilience

The resilience of a rubber may be considered as the energy recoverable after deformation. A simple example is where a metal ball is dropped onto a slab of rubber from a height H. If the ball rebounds to a height $0.7H$ it would be said that the resilience of the rubber is 70%. Energy not recovered is converted to heat. Since no rubber is 100% resilient it follows that if a piece of rubber is constantly flexed or otherwise deformed there is a build up of heat which, unless dissipated, causes a temperature rise and possibly failure of the part. For this reason such components as sidewalls of lorry tyres need to be made from high resilience compounds.

Above T_g, resilience increases with temperature up to about $(T_g + 100)$°C. Since the rate of change of resilience with temperature is similar for many, but not all, rubbers, this also tends to mean that the lower the T_g of the rubber the higher the resilience at normal ambient temperature (*Figure 2.8*).

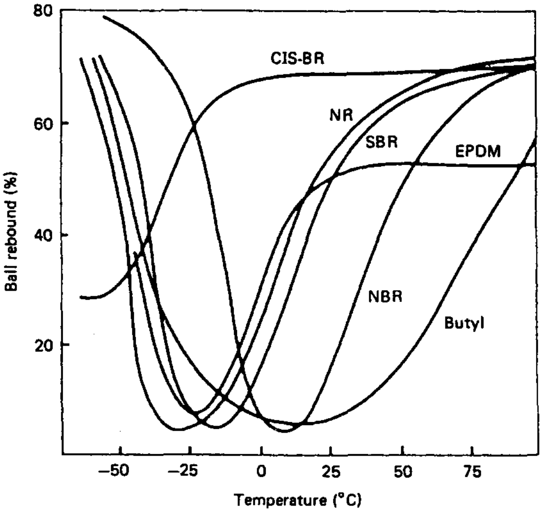


Figure 2.8 Resilience (ball rebound) vs. temperature relationships for butyl rubber compared with other major rubbers. (Reproduced by permission from Gunter, *Developments in Rubber Technology*, 2nd edition, Applied Science)

2.2.7 Creep, set and dynamic properties

When a polymer is subjected to a tensile stress the total deformation may be considered as the sum of three components:

(1) an instant deformation due to bond bending and stretching (D_{OE}) similar to that of ordinary elastic bending in metals;
(2) chain uncoiling which is not instantaneous and leads to high elastic deformation (D_{HE}); and
(3) chain slippage leading to viscous deformation (D_{visc}).

Hence,

$$D_{total} = D_{OE} + D_{HE} + D_{visc}$$

These various responses and the total response are shown in *Figures 2.9* and *2.10*.

The relative proportions of each component vary with the type of polymer, but with a vulcanised rubber the D_{HE} component predominates. Even so there will be some irreversible *creep* which will progress with time. Similarly, if a piece of rubber is held under constant strain, the stress required to maintain this strain decreases with time. This phenomenon is known as *stress decay*.

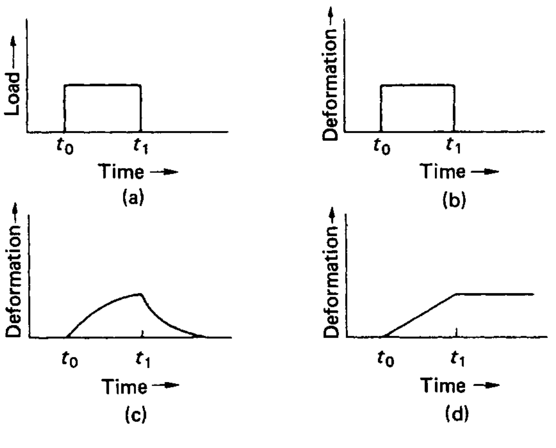


Figure 2.9 Types of deformational response as a result of a fixed load being imposed between times t_0 and t_1 (a). (b) Ordinary elastic material. (c) Highly elastic material. (d) Viscous material. (Reproduced by permission from Brydson, *Rubbery Materials and their Compounds*, Elsevier/Applied Science)

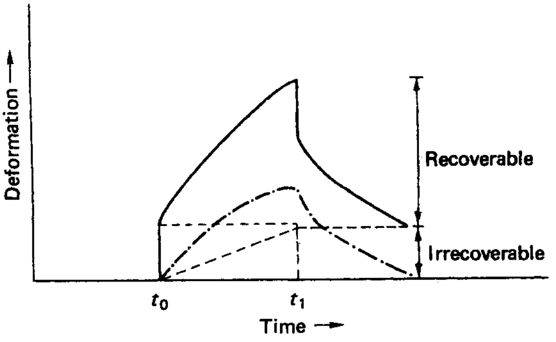


Figure 2.10 Deformation vs. time curve for material showing substantial ordinary elastic, high elastic and viscous components of deformation. (Reproduced by permission from Brydson, *Rubbery Materials and their Compounds*, Elsevier/Applied Science)

Total deformations in other strain modes have similar relationships and, in practice, since rubber is more often used in compression and shear, these modes can be more important. In particular it is important that, when rubber is being used as a compression seal, the stress decay is not so great that seal fails. Also of concern is the *compression set*, i.e. the irrecoverable component of a compressive deformation.

Because D_{HE} is not instantaneous, strain will lag behind stress in a cyclic deformation. The mathematics are similar to those in a.c. theory and at this point just two points will be made.

(1) If the stressing frequency is progressively increased there will be less time for the strain to reach equilibrium so that the response in each cycle will decrease. This will effectively increase the modulus.
(2) By paying regard to the natural vibrating frequency of a rubber component it is possible to isolate vibrations. This technique is useful, for example, for protecting delicate instruments from machinery vibrations and buildings from earthquakes.

2.2.8 Non-rubbery properties of technical importance

Clearly there are many properties other than stress–strain properties that are often important in the use of a rubbery material. These may include heat resistance, oil resistance, electrical insulation behaviour, general chemical resistance, permeability to gases, colour and, always importantly, cost. In this subsection brief consideration is given to the first two of these properties.

In the case of a cross-linked (vulcanised) rubber, the heat resistance will be mainly concerned with the degradation in structure that occurs at elevated temperatures. This will depend on both the chemical structure of the raw polymer *and* on the vulcanising system used. Therefore it is important to understand that any statement concerning the relative heat resistance of various rubbery polymers needs to be given some latitude depending on the vulcanising system employed.

In the case of thermoplastic rubbers the upper service limit is determined by the softening point at which the material loses its rubberiness and begins to flow.

In many applications rubbery components are often in contact with hydrocarbon oils. As a general rule the hydrocarbon rubbers (which form the bulk of the market) have less oil resistance than do non-hydrocarbon rubbers.

2.2.8.1 Classification system for vulcanised rubbers

The American Society for Testing and Materials (ASTM) and the Society of Automotive Engineers (SAE) have issued a report entitled *Classification System for Rubber Materials for Automotive Applications* (SAE J200/ASTM D2000); BS 5176 issued by the British Standards Institution is almost identical. The classifications do not at present apply to thermoplastic rubbers.

The classification provides basic classification codes based on test values obtained in standard laboratory tests. Each code is made up of two parts: the first part based on *basic requirements*; the second on *supplementary* or *suffix requirements*. These provide, for a given specification, a *line call-out*.

The first part of the line call-out results from the basic requirements and consists of:

(1) a letter 'M' to indicate that SI units are being used (this may be followed or preceded, according to which specifi-

cation is being used, by a number indicating supplementary requirements);

(2) a letter indicating the *type*—assigned on the level of heat resistance required;

(3) a letter indicating the *class*—assigned on the level of mineral oil (hydrocarbon oil) resistance required;

(4) a single figure indicating the hardness range; and

(5) a double figure indicating the tensile strength.

The *type* is based on changes in tensile strength of not more than 30%, an elongation change of −50% maximum and a hardness change of not more than 15 points. The temperatures at which the materials are tested for determining type are shown in *Table 2.4*.

The *class* is based on the resistance of material to swelling in ASTM Oil No. 3 (or in the case of BS 5176, Oil No. 3 of BS 903). In test the immersion time shall be 70 h and the oil temperature shall be the type temperature or 150°C, whichever is the lower. Limits of swelling are listed in *Table 2.5*.

The basic mechanical properties are indicated by a single figure for hardness followed by a double figure for tensile strength. Thus the set '810' indicates a Shore hardness (a commonly used hardness scale) of 80 ± 5 and a tensile strength of 10 MPa minimum. (The BS specification uses IRHD and with slightly different tolerance limits of +5 to −4.)

An example of an ASTM D2000 line call-out is: M6 HK 810 A1-11 B38 EF31. This indicates the use of SI units, an elastomer of type H (resistant to 250°C) and of class K (10% swell), a hardness of 80 ± 5 and a minimum tensile strength of 10 MPa. The number '6' after 'M' provides a key to supplementary suffix requirements which are detailed in the second part of the call-out (A1-11 B38 EF31). Interpretation of suffix requirements requires reference to details in the appropriate standards, but it may be noted that in the above example call-out these refer to particular heat resistance, compression set and fuel resistance requirements.

Figure 2.11 shows in graph form the relative positions of selected rubbers in respect of type and class. *Table 2.6* provides a summary of properties of most commercial rubbery materials.

Table 2.4 Classification of rubber by type—assigned according to the level of heat resistance

Type	*Test temperature (°C)*	*Type*	*Test temperature (°C)*
A	70	F	200
B	100	G	225
C	125	H	250
D	150	J	275
E	175		

Table 2.5 Classification of rubber by class—assigned according to the level of resistance to mineral oil

Class	*Maximum volume swell (%)*	*Class*	*Maximum volume swell (%)*
A	No requirement	F	60
B	140	G	40
C	100	H	30
D	100	J	20
E	80	K	10

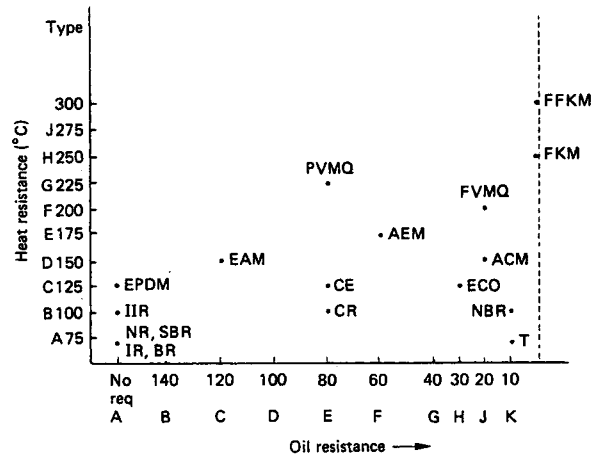


Figure 2.11 SAE J200 ASTM D2000 classification for cross-linked elastomers. Whilst specific points are shown for clarity, there will be a spread for each rubber, the exact position depending on the grade of polymer and compounding formula used. (Reproduced by permission from Brydson, *Rubbery Materials and their Compounds*, Elsevier/Applied Science)

2.3 Survey of commercial rubbery materials

There are about 30 types of rubbery material on the market, although just two (natural rubber and styrene–butadiene rubber) account for about 70% of total tonnage and three for about 80%. These materials are listed together with their recommended ISO/ASTM abbreviations in *Table 2.7*.

In addition to the nomenclature based on ISO and ASTM recommendations, several other abbreviations are widely used. Those most likely to be encountered are listed in *Table 2.8*.

During World War II the US Government introduced a system of nomenclature which continued in use, at least partially, until the 1950s and is used in many publications of the period (see *Table 2.8*).

Rubber usage continues to be dominated by the tyre industry, which consumes about two-thirds of total new rubber each year. Hence it is not surprising that those rubbers widely used in the industry tend to dominate the market, not simply because they are used in tyres but also because the large-scale production tends to give these materials a competitive price structure. The term *tyre rubbers* is given to these materials.

In the non-tyre industry, tyre rubbers are used alongside other rubbers and four such materials are used in this area in quantities similar to those of the tyre rubbers. These materials are commonly known as *special purpose rubbers*, although in reality they may be used in a wider range of applications than some of the tyre rubbers.

There are a number of applications, particularly in engineering, where more severe demands are put on the rubber than can be met with the tyre rubbers and special purpose rubbers, particularly in terms of heat and oil resistance. The term *speciality rubber* is commonly used for these materials. Some extremely specialised materials, available only in small quantities, and at high cost are sometimes separately classed as *exotics*.

Table 2.6 Summary of properties of principal rubbery materials*

	ASTM D1418 class							
Property	*NR*	*SBR*	*BR*	*IR*	*EPDM*	*IIR*	*CR*	*NBR*
ASTM D2000 rating	AA	AA/BA	AA/BA	AA	CA	BA, AA	BC, BE	BF, BG, BK, CM
Specific gravity	0.92	0.94	0.93	0.93	0.86	0.92	1.23	1.00
Tensile strength (gum) (MPa (psi))	>20.7 (3000)	<6.9 (1000)	<6.9 (1000)	>17.2 (2500)	<6.9 (1000)	>10.3 (1500)	>20.7 (3000)	<6.9 (1000)
Tensile strength (black loaded) (MPa (psi))	>20.7 (3000)	>13.8 (2000)	>20.7 (3000)	>20.7 (3000)	>20.7 (3000)	>13.8 (2000)	>20.7 (3000)	>13.8 (2000)
Tear resistance	Good–v.good	Fair	Fair	Good–v. good	Fair–good	Good	Good	Fair
Abrasion resistance	Excellent	Good–excellent	V. good	V. good	Good–excellent	Good	Excellent	Good
Compression set	Good	Good	Good	V. good	Good	Fair	Fair–good	Good
Rebound (cold)	Excellent	Good	Excellent	Excellent	V. good	Poor	V. good	Good
Rebound (hot)	Excellent	Good	Excellent	Excellent	Excellent	Good	V. good	Good
Dielectric strength	Excellent	Excellent	V. good	V. good	Excellent	Excellent	Good	Poor
Electrical insulation	Good–excellent	Good–excellent	V. good	V. good	Excellent	Good–excellent	Fair–good	Poor
Environmental resistance								
Oxidation	Good	Fair	Fair	Fair	Excellent	Excellent	V. good–excellent	Good
Ozone	Poor–fair	Poor	Poor	Poor	Outstanding	Excellent	V.good	Fair
Sunlight	Poor	Poor	Poor	Poor	Outstanding	V. good	V.good	Poor
Heat ageing (upper limit at constant service)(°C)	85	90	90	85	145	120	95	120
Low-temperature flexibility	Excellent	V. good	Excellent	Excellent	Excellent	Good	Good	Fair–good
Fluid resistance								
Gas permeability	Fair	Fair	Fair	Fair	Fair	V. low	Low	V. low
Acid resistance (dilute)	Good	Good	Good	Good	Excellent	Excellent	Excellent	Good
Acid resistance (conc.)	Good	Good	Good	Good	Excellent	V. good	Good	Good
Swelling								
In aliphatic hydrocarbons	Poor	Poor	Poor	Poor	Poor	Poor	Fair–good	Excellent
In aromatic hydrocarbons	Poor	Poor	Poor	Poor	Poor	Poor	Fair	Good
In oxygenated solvents	Fair–good	Good	Good	Fair–good	Good	Good	Poor	Poor

	ASTM D1418 class							
Property	*ACM*	*ECO/CO*	*AEM*	*CSM*	*FKM*	*PVMQ*	*FVMQ*	*YBPO*
ASTM D2000 rating	DJ	CM	EF	CE	HK	GE	FJ	Not assigned
Specific gravity	1.10	1.27–1.36	1.08–1.12	1.12–1.28	1.85	1.06–1.42 (compound)	1.41 (compound)	1.17–1.25
Tensile strength (gum) (MPa (psi))	—	<6.9 (1000)	—	>17.2 (2500)	>12.4 (1800)	—	—	25–45 (3500–6500)
Tensile strength (black loaded) (MPa (psi))	>13.8 (2000)	17.2 (2000)	>17.2 (2500)	>20.7 (3000)	>13.8 (2000)	5–10 (725–1450)	>6.9 (1000)	—
Tear resistance	Fair	Fair–good	Good	Fair	Fair	Poor	Poor	Outstanding
Abrasion resistance	Fair	Fair–good	Good	Excellent	Good	Poor	Poor	Outstanding
Compression set	V. good	Poor	Good	Fair	Fair–good	Poor–fair	Poor–fair	Fair–poor
Rebound (cold)	Low	Good	Poor	Fair–good	Fair–good	Poor	Poor	Fair–v. good
Rebound (hot)	—	Good	Fair	Good	Good	—	—	Good–exclent
Dielectric strength	Good	Good	Good	V. good–excellent	Good	Good	Good	Fair–good
Electrical insulation	Fair	Good	Fair–good	Good	Fair–good	Excellent	V. good	Fair–good
Environmental resistance								
Oxidation	Good	Good	Excellent	Excellent	Outstanding	Excellent	Excellent	Excellent
Ozone	Excellent	Excellent	Outstanding	Outstanding	Outstanding	Excellent	Excellent	Excellent
Sunlight	Good	Good	Outstanding	Outstanding	Outstanding	Excellent	Excellent	V. good
Heat ageing (upper limit at constant service)(°C)	177	135	170	135	205	235	230	100–110
Low-temperature flexibility	Fair–v. good	Good–v. good	Good	Good	Fair–good	Excellent	Excellent	Excellent
Fluid resistance								
Gas permeability	Fair	Low	V.low	V. low	V. low	High	High	Low
Acid resistance (dilute)	Fair	Fair–good	Good	Excellent	Excellent	Excellent	Excellent	Excellent
Acid resistance (conc.)	Poor	Fair	Poor	V. good	Excellent	Fair	Good	V. good
Swelling								
In aliphatic hydrocarbons	Excellent	Excellent	Good	Good	Excellent	Poor	Excellent	Excellent
In aromatic hydrocarbons	Fair–poor	Good	Fair	Fair	Excellent	Poor	Good–excellent	Good
In oxygenated solvents	Poor	Poor	Poor	Poor	Poor	Fair	Good–excellent	Fair

* Reproduced by permission from Brydson, *Rubbery Materials and Their Compounds*, Elsevier/Applied Science.

Table 2.7 Standard abbreviations for rubbery materials (based on ISO Recommendation and ASTM D 1418)

ABR	Acrylate–butadiene rubber
ACM	Copolymers of ethyl or other acrylates and a small amount of a monomer which facilitates vulcanisation
AECO	Terpolymers of allyl glycidyl ether, ethylene oxide and epichlorohydrin
AEM	Copolymers of ethyl or other acrylate and ethylene
AFMU	Terpolymer of tetrafluoroethylene, trifluoronitrosomethane and nitrosoperfluorobutyric acid
ANM	Copolymers of ethyl or other acrylate and acrylonitrile
AU	Polyester urethanes
BIIR	Bromo-isobutene-isoprene rubber (brominated butyl rubber)
BR	Butadiene rubber
CFM	Polychlorotrifluoroethylene
CIIR	Chloro-isobutene-isoprene rubber (chlorinated butyl rubber)
CM	Chlorinated polyethylene
CO	Epichlorhydrin rubber
CR	Chloroprene rubber
CSM	Chlorosulphonated polyethylene
ECO	Ethylene oxide and epichlorhydrin copolymer
EAM	Ethylene–vinyl acetate copolymer
EPDM	Terpolymer of ethylene, propylene and a diene with the residual unsaturated portion of the diene in the side-chain
EPM	Ethylene–propylene copolymer
EU	Polyether urethanes
FFKM	Perfluoro rubbers of the polymethylene type having all substituent groups on the polymer chain either fluoro, perfluoroalkyl or perfluoroalkoxy groups
FKM	Fluoro rubber of the polymethylene type having substituient fluoro and perfluoroalkoxy groups on the main chain
FVMQ	Silicone rubber having fluorine, vinyl and methyl substituient groups on the polymer chain
GPO	Polypropylene oxide rubbers
IIR	Isobutene–isoprene rubber (butyl rubber)
IM	Polyisobutene
IR	Isoprene rubber (synthetic)
MQ	Silicone rubbers having only methyl substituent groups on the polymer chain
NBR	Nitrile–butadiene rubber (nitrile rubber)
NIR	Nitrile–isoprene rubber
NR	Natural rubber
PBR	Pyridine–butadiene rubber
PMQ	Silicone rubbers having both methyl and phenyl groups on the polymer chain
PNR	Polynorborene rubber
PSBR	Pyridine–styrene–butadiene rubber
PVMQ	Silicone rubbers having methyl, phenyl and vinyl substituent groups on the polymer chain
Q	Rubbers having silicon in the polymer chain
SBR	Styrene–butadiene rubber
T	Rubbers having sulphur in the polymer chain (excluding copolymers based on CR)
VMQ	Silicone rubber having both methyl and vinyl substituent groups in the polymer chain
XNBR	Carboxylic–nitrile butadiene rubber (carboxynitrile rubber)
XSBR	Carboxylic–styrene butadiene rubber
Y	Prefix indicating thermoplastic rubber
YBPO	Thermoplastic block polyether–polyester rubbers

Two further groups of rubbery materials which have a quite different processing technology to the above are the *thermoplastic rubbers* which are processed on equipment used in thermoplastics processing and the *liquid processing rubbers* such as the cast polyurethanes and the liquid silicones which have their own specialist technology.

2.3.1 Tyre rubbers

This term is used for three diene rubbers widely used in the tyre industry. Although capable of making products with good performance in respect of toughness, strength, abrasion and tear resistance these materials have a lower resistance to hydrocarbon oils, oxygen, ozone, heat and light than do most of the rubbers described in the following sections.

All of the tyre rubbers are usually vulcanised using conventional sulphur based systems.

Table 2.8 Miscellaneous abbreviations used for rubbery materials

ENR		Epoxidised natural rubber
EPR		Ethylene–propylene rubbers (either EPM or EPDM)
EVA		Ethylene–vinyl acetate copolymers (instead of EAM)
SBS		Styrene–butadiene–styrene triblock copolymer
SEBS		Hydrogenated SBS
SIR		Standard Indonesian rubber
SIS		Styrene–isoprene–styrene triblock copolymer
SMR		Standard Malaysian rubber
US Government nomenclature		
GR-A	Government Rubber	Acrylonitrile (modern equivalent NBR)
GR-I	Government Rubber	Isobutylene (IIR)
GR-M	Government Rubber	Monovinyl acetylene (CR)
GR-P	Government Rubber	Polysulphide (T)
GR-S	Government Rubber	Styrene (SBR)

2.3.1.1 Natural rubber

Natural rubber (NR) has about 40% of the total new rubber market in tonnage terms. Although rubbery materials have been obtained from the latices (plural of latex) of some 500 plants, nearly all commercial rubber comes from *Hevea brasiliensis*, originally from Brazil but now grown widely in South-East Asia, West Africa and the Indian subcontinent. The material is obtained by coagulating the latex which has been tapped from the bark of the tree. The raw polymer has a high molecular weight (>1 million) and, because of entanglements, is quite elastic and incapable of shaping by melt processing. It is therefore necessary to subject the rubber to an intensive shearing process (mastication) to break down the molecular weight to a state in which the material may be blended with additives, shaped and then vulcanised. Newer grades may be treated at the latex stage to produce raw material which needs little or no mastication (known as 'constant viscosity (CV) rubbers').

Unfilled vulcanisates (gum stocks) are highly elastic and resilient and, although this may be reduced by incorporation of fillers, NR vulcanisates tend to be better than comparable filled vulcanisates from most synthetic rubbers. On ageing, the rubber tends to soften.

Because of its high resilience NR is widely used in tyre sidewalls and, more generally, in large tyres such as truck tyres. It is also an engineering rubber, where it is used, for example, in engine mountings and bridge bearings. Unlike other tyre rubbers, gum stocks show high vulcanisate strengths and this leads to many uses in stationery and medical applications.

The two major disadvantages of NR are its poor resistance to hydrocarbon oils (vulcanisates swelling extensively) and moderate heat resistance. However, penetration by oil and by oxygen (the main agent for limiting heat resistance) is slow so that thick articles or articles in which a large portion of the surface is in contact with metal and thus not accessible to oil and/or oxygen may be satisfactorily used at high temperatures or in oily environments for long periods of time.

2.3.1.2 Styrene–butadiene rubber

For many years styrene–butadiene rubber (SBR) was the leading rubber in tonnage terms, NR recovering its pre-

eminence about 1980. Currently SBR has about 30% of the market. It differs from NR in the following ways.

(1) It does not break down on mastication and so has to be supplied at the correct molecular weight. Unvulcanised compounds also have inferior strength (green strength) and lack natural tack (a property useful when assembling or plying up pieces of rubber to make an article).
(2) Carbon black reinforcement is necessary for reasonable strength.
(3) Vulcanisation is slower than with NR with similar vulcanising (curing) systems.
(4) Abrasion resistance of reinforced compounds is generally better than with natural rubber. Wet grip and rolling resistance of tyres is also better.
(5) It hardens, rather than softens, on oxidation.

Mainly used for tyres, SBR is also used in such areas as mechanical goods, shoe soling and carpet underlay.

2.3.1.3 Polybutadiene rubber

Polybutadiene rubber (BR) is superior to both NR and SBR in terms of:

(1) abrasion resistance and groove cracking of tyres;
(2) low-temperature flexibility;
(3) high resilience at low deformations;
(4) heat ageing resistance;
(5) ozone resistance (under both static and dynamic conditions); and
(6) ability to accept higher levels of fillers and oil with less deterioration in properties.

It does, however, exhibit the following limitations:

(1) poor tack;
(2) poor road grip of tyre treads; and
(3) poor tear and tensile strength (compared with NR and SBR).

With regard to (3), in some ways BR vulcanisates may be considered as brittle with the possibility of products being subject to chunking and gouging, that is pieces of rubber breaking off when operating under severe conditions.

For these reasons BR is seldom used alone but mainly as a blending material (e.g. with SBR in passenger-car tyres and NR in truck tyres). It is also widely used as a toughening rubber in the manufacture of high-impact polystyrene. A small application is for very highly resilient solid playballs. BR has about 10% of the new rubber market.

2.3.2 Special purpose rubbers

This term is something of a misnomer but is commonly employed for those rubbers that find fairly widespread use but which are not normally used in any quantity as tyre rubbers.

2.3.2.1 Ethylene–propylene–diene terpolymers

These materials are now widely used in the general rubber-goods field. They have the best heat resistance of the tyre and special purpose rubbers as well as good resistance to oxygen, ozone and to light. As with the diene tyre rubbers (BR, NR and SBR) they do not have good resistance to hydrocarbon oils. Unlike NR, but like SBR, they require reinforcement with fine fillers in order to obtain good strength and abrasion resistance.

Ethylene–propylene–diene (EPDM) polymers contain just a small amount of some diene component which allows the rubber to be vulcanised using conventional sulphur based curing systems.

2.3.2.2 Ethylene–propylene copolymers

With properties similar to the above, the absence of a diene component in ethylene–propylene (EPM) copolymers prevents the use of conventional sulphur cures and require such materials as peroxides for vulcanisation. These rubbers therefore have limited use compared with EPDM.

2.3.2.3 Isobutylene–isoprene rubbers (butyl rubbers)

Isobutylene–isoprene rubbers (butyl rubbers; IIR) have the following particular characteristics:

(1) very low air permeability for a rubber (but not as low as many plastics materials);
(2) low resilience over a wide temperature range, although resilience does increase with temperature; and
(3) low unsaturation (i.e. low level of double bonds in the polymer) giving good resistance to oxygen, ozone and weathering, although not so good as with EPM and EPDM.

As with EPDM, a small amount of a diene component (from isoprene) allows conventional sulphur vulcanisation.

Comparatively recent developments are the brominated and chlorinated butyl rubbers (BIIR and CIIR). They are faster curing and capable of vulcanisation bonding. The main uses for IIR and derivatives are tyre inner tubes and inner linings, tyre curing bags, sealants, pipe-wrapping systems, pharmaceutical bottle closures, acid-resistant tank linings, reservoir roofing membranes, pond liners, and even chewing gum.

2.3.2.4 Polychloroprene rubbers

Available since the early 1930s polychloroprene (CR) rubbers are well known under the tradenames 'Neoprene', 'Baypren' and 'Butachlor'. They are well established as engineering rubbers of moderate heat, oxygen and ozone resistance with useful fire retardancy characteristics and better hydrocarbon oil resistance than any of the hydrocarbon based rubbers described earlier in this section.

They cannot be vulcanised using conventional sulphur based systems and require their own specialist curing technology.

2.3.2.5 Butadiene–acrylonitrile rubbers (nitrile rubbers)

The principal use for butadiene–acrylonitrile rubbers (NBR) is where oil, grease and, particularly, petrol resistance is required. Recently, more severe demands for under-the-bonnet applications have required NBR to be replaced by some of the speciality rubbers.

Nitrile rubbers may be vulcanised using conventional sulphur based systems.

2.3.2.6 Polyisoprene rubber

Polyisoprene rubber (IR) is a synthetically produced version of NR. Although generally purer than the natural product the polymer structure is less perfect. This means that the mechanical properties are generally not so good but, because of the lower level of impurities, IR finds use in pharmaceutical goods and as the raw material for producing chlorinated and isomerised rubber for the surface-coating industry. Other specialised uses include golf-ball thread (because of the low modulus) and some engineering applications where the low compression set (compared with NR) is an important prerequisite.

At one time IR was extensively used in tyre blends, but in recent years this application has proved uneconomical (rather than technically unsatisfactory).

2.3.3 Speciality rubbers

2.3.3.1 Acrylic rubbers

Acrylic rubbers (ACM) were developed for use where it is necessary to operate at higher temperatures than is possible for nitrile rubbers or where oils contain sulphur bearing additives which tend to react with nitrile rubber. They are also superior in heat resistance to all the rubbers considered above.

2.3.3.2 Fluoro rubbers

Many types of fluoro rubber are now available but, in general, they are used where a high level of oil resistance coupled with a high level of heat resistance is required.

2.3.3.3 Silicone rubbers

These materials are used because of their excellent heat resistance and excellent electrical insulation properties, their physiological inertness, their high permeability and low compression set. Besides conventional materials, special grades include liquid silicones rubbers and room temperature vulcanising systems that may be used for flexible mould-making, for seals and for caulking applications.

2.3.3.4 Polysulphides

Polysulphides (T) have been in production for longer than any synthetic rubber, having been introduced in the 1920s. They have poor heat resistance but outstanding resistance to hydrocarbon oils.

2.3.3.5 Epichlorhydrin rubbers

Epichlorhydrin (CO and ECO) rubbers have excellent resistance to ozone and weathering, an even lower gas permeability than butyl rubber, and good resistance to both heat and swelling in hydrocarbon oils. They have replaced nitrile rubber in some applications where specifications have become too stringent for the latter.

2.3.3.6 Polynorbornenes

Polynorbornenes (PNR) are hydrocarbon rubbers, of interest because they can be blended with large amounts of plasticiser and thus are available in very soft forms. They can also be formulated with interesting damping characteristics.

2.3.3.7 Chlorinated polyethylene and chlorosulphonated polyethylene

Chlorinated polyethylene (CM) and chlorosulphonated polyethylene (CSM) have good resistance to heat, ozone, weathering and a wide range of chemicals. Their main uses are in coated fabric and sheet form and for wire covering.

2.3.3.8 Ethylene–acrylic rubbers

Ethylene–acrylic (AEM) rubbers have intermediate heat resistance to more expensive rubbers and also may be formulated to have good flame retardance coupled with low emission of smoke and toxic gases under fire conditions. Uses are split between cable and automotive applications.

2.3.4 Thermoplastic rubbers

All the materials discussed in this subsection can be processed on conventional thermoplastic processing machinery such as injection-moulding machines.

2.3.4.1 Styrene–butadiene–styrene triblock rubbers

The high level of creep with styrene–butadiene–styrene (SBS) triblock rubbers limits their use in conventional rubber applications. They are important as bitumen additives, in pressure-sensitive and contact adhesives and for plastics modification. They are also used in shoe soling. Related SIS rubbers using isoprene instead of butadiene are also useful in adhesives.

2.3.4.2 Thermoplastic polyester rubbers and similar materials

These materials have better oil resistance and heat resistance (both in terms of thermal stability and heat softening resistance) than SBS materials. They are also available in a range of hardness from a stiff rubber to that of a more leather-like plastics material. They are important specialised engineering rubbers. Somewhat similar in properties are the thermoplastic polyurethanes and the thermoplastic polyamide rubbers.

2.3.4.3 Thermoplastic polyolefin rubbers

These materials are blends of EPDM with polypropylene, and sometimes also polyethylene. Being hydrocarbons they have limited oil resistance. They are of extensive interest in automotive applications, window profiles, cable insulation and footwear. Although they have somewhat limited levels of high elastic recovery after deformation, there have been considerable improvements with these properties in recent years.

2.3.5 Liquid processing rubbers

Both polyurethane and silicone rubbers are frequently processed in the liquid (low-molecular-weight state). Their processing technology is, therefore, highly specialist. Polyurethanes made using these techniques may have, if suitably formulated, outstanding oil resistance and abrasion resistance coupled in many cases with high values of tensile strength. Silicones tend to retain the properties mentioned in Section 2.3.3.3.

2.3.6 Exotics

The term 'exotics' is applied to highly specialised materials often only produced to specific order. They include the fluoropolyphosphazenes which have a very wide operating range (−65 to +175°C) coupled with good oil resistance and the totally non-inflammable carboxy–nitroso rubbers (AFMU).

Rubbery materials are discussed in greater detail in *Rubbery Materials and their Compounds*[1] with more recent developments discussed in more detail in *Handbook of Elastomers*.[2]

2.4 General properties of plastics

Since plastics may be regarded as those polymers that are not rubbers, fibres, adhesives or surface coatings, the range of properties possible is very wide. Some materials such as low-density polyethylene and plasticised PVC are flexible; the polyurethane foams are flexible and, in some cases, rubbery;

glass-reinforced polyester laminates are hard and rigid; whilst the nylons are less rigid but tough. Some plastics dissolve in water, others have excellent resistance to a wide range of chemicals. Other properties such as electrical-insulation characteristics and heat resistance also vary widely.

Since this handbook is concerned with engineering, this section is restricted to the consideration of certain selected properties of particular importance when plastics are used in engineering applications. The next section then provides a brief review of the main types of plastics material.

The properties considered here are:

(1) softening point (T_g and T_m);
(2) heat resistance;
(3) fire retardancy;
(4) toughness;
(5) creep resistance; and
(6) solvent resistance.

2.4.1 Softening point

The maximum temperature that may be used with a given plastics material depends on two independent factors:

(1) the softening behaviour of the polymer; and
(2) the thermal stability of the polymer, particularly in air.

Whilst the softening behaviour will put a short-term limit on the maximum temperature, the thermal stability is more time dependent.

The softening point of a plastics material is an arbitrary property, being the point at which the modulus of the material drops to some arbitrary, and usually unmeasured, value. Two tests are now widely used.

The Vicat softening point test A sample of the material is heated at a specified rate of temperature increase and the temperature is noted at which a needle of specified dimensions indents into the material a specified distance under a specified load. Most commonly a load of 10 N is applied, the needle indentor cross-section is 1 mm^2, the specified penetration distance is 1 mm and the rate of temperature rise is 50°C per hour. (The relevant standards are ISO 306; BS 2782: Method 120; ASTM D1525 and DIN 53460.)

The deflection temperature under load test Also widely known by its earlier name of the *heat distortion temperature test*, the temperature is noted at which a bar of material, subjected to a three-point bending stress, is deformed a specified amount. The load F applied to the sample will vary with the thickness t and width w of the sample and is determined by the maximum stress specified at the midpoint of the beam P which may be either 0.45 MPa or 1.82 MPa. (The formula used for the calculation is $F = 2Pwt^2/3L$, where L is the distance between the outer supports.) (The relevant standards are ISO 75, BS 2782: Method 121, ASTM D648 and DIN 53461.)

It can be seen from *Figure 2.2* that, in the case of amorphous polymers such as polystyrene and polymethyl methacrylate, the modulus will eventually reach a temperature where the modulus drops catastrophically. In these circumstances the softening points are very similar whatever test method is used and are all quite close to T_g (*Table 2.9*).

In the case of crystalline polymers, modulus does not change so sharply with temperature and a wide spread of softening point values are obtained. In this case Vicat temperatures are often quite close to T_m (*Table 2.10*).

Table 2.9 Glass transition temperature (T_g), Vicat softening point (VSP) and deflection temperature under load (DTL) of selected amorphous plastics

Plastic	T_g (°C)	*VSP* (°C)	*DTL (at 1.82 MPa)* (°C)
Polycarbonate*	149	155	127–135
Polymethyl methacrylate	104	114	85–100
Polystyrene	90–100	90	80
Polysulphone	190	—	174
Polyethersulphone	230	226	203
UPVC	80	73–77	76–80

* Polycarbonate is very slightly crystalline.

Table 2.10 Relationship between T_g, T_m and softening points for selected crystalline plastics

Plastic	T_g (°C)	T_m (°C)	*VSP* (°C)	*DTL (at 1.82 MPa)* (°C)
Nylon 6	50	215	—	80
Nylon 66	60	264	240	80
Polyacetal (copolymer)	−13	163	162	110
PBT	22–43	225	169–180	65

Many of the so-called *engineering thermoplastics* (see below) are supplied in fibre-reinforced form. It is interesting to note here that fibre reinforcement has a greater effect on the softening point of crystalline than on amorphous polymers (*Figure 2.12*).

It is of further interest to note that with such polymers the deflection temperature of reinforced polymers is often close to T_m, whilst that of the unreinforced materials is closer to T_g (*Table 2.11*).

2.4.2 Thermal stability

As with assessment of softening point, the assessment of thermal stability is also somewhat arbitrary. One method that

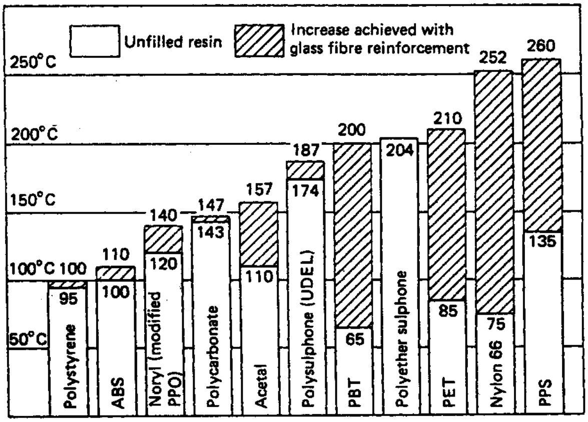


Figure 2.12 Heat deflection temperatures under a load of 1.82 MPa for selected polymers. Note that incorporation of glass fibre has a much greater effect with crystalline polymers than with amorphous ones. (Reproduced by permission from Whelan and Craft, *British Plastics and Rubber*, 29 (Nov. 1982))

Table 2.11 Comparison of T_g, T_m and heat deflection temperatures of polymers with and without glass reinforcement*

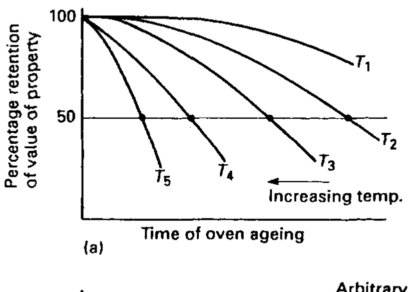
	T_g (°C)	T_m (°C)	*Heat deflection temperature* (°C) *Unfilled*	*Filled*
Polyether ether ketone	143	334	150	315
Polyether ketone	165	365	165	340
Polyphenylene sulphide	85	285	135	260
Polyethylene terephthalate	70	255	85	210
Polybutylene terephthalate	22–43	225	54	210

* Reproduced by permission from Brydson, *Plastics Materials*, 5th edition, Butterworth-Heinemann.

has become increasingly used in recent years is the test leading to the Continuous Use Temperature Index of the Underwriters Laboratories.

In order to obtain this rating a large number of samples are subjected to oven ageing at a variety of temperatures for periods up to 1 year. In this time samples are withdrawn and tested and values of the property tested are plotted against time. The time taken for the property to lose half its value is noted for each temperature tested. This time is, quite arbitrarily, referred to as the *failure time*. A plot of the logarithm of the failure time against 1/*K* (where *K* is the temperature in kelvin) is made and the temperature corresponding to a failure time of 10 000 h is referred to as the *temperature index* (*Figure 2.13*).

Clearly, the temperature index will depend on the property being measured. Amongst those properties widely used are tensile strength (where the results are often quoted as 'mechanical, without impact'), Izod impact strength and electrical properties. Results also depend on sample thickness. Some representative data are given in *Table 2.12*.



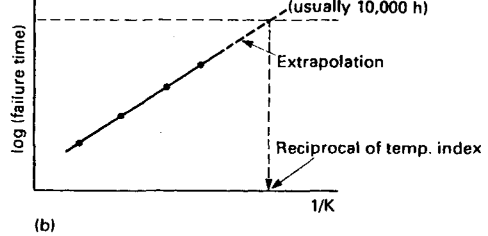


Figure 2.13 Method for determining the UL temperature index. (Reproduced by permission from Brydson, *Handbook for Plastics Processors*, Butterworth-Heinemann)

Table 2.12 Continuous Use Temperature Index (mechanical, without impact unless otherwise stated)*†

Polymer	*Temperature* (°C)
ABS	60–80
Nylons	75
Polyacetal (homopolymer)	90
Polyacetal (copolymer)	90
Polybutylene terephthalate	105–140
Polyethylene terephthalate	140 (glass reinforced)
Polyarylate (Ardel D-100)	130
Polysulphones	160
Polyetherimide	170
Polyethersulphone	180
Polyphenylene sulphide‡	200
Aromatic polyester (Ekkcel I-2000)	220
Liquid crystal polyester	220
Polyether ether ketone	240

* Reproduced by permission from Brydson, *Handbook for Plastics Processors*, Butterworth-Heinemann.
† The data given in this table have been obtained using procedures related to the UL Temperature Index method. The data given here do not, however, necessarily imply that the Underwriters Laboratories have awarded the temperature ratings quoted here either for the class of materials or for specific grades. Unless specified the figures are for standard unreinforced polymers.
‡ Glass reinforced, mechanical with impact.

2.4.3 Flame retardancy

For many applications resistance to burning is an important requirement. Whilst in many cases it is the fire reistance of a total structure that is of prime concern, the inherent flame-retardancy properties of the plastics materials used are also important. Two tests, described in the next two subsections, are now widely used for this purpose. It is also to be noted that in a fire more deaths occur through inhalation of smoke and toxic gases than through direct burning. For this reason assessments should also be made of smoke evolution and of toxicity of fire products.

2.4.3.1 Limiting oxygen index

The 'limiting oxygen index' (LOI) is the minimum concentration of oxygen, expressed as a percentage, in a slowly rising stream of a mixture of oxygen and nitrogen which will support combustion. The percentage of oxygen in air is 20.9%, so that polymers with a higher figure than this will not burn easily, if at all. Typical LOI figures are given in *Table 2.13*.

2.4.3.2 Underwriters Laboratory UL 94 Vertical Burning Test

Where plastics materials exhibit a measure of flame retardancy it is more common to classify them according to the Underwriters Laboratory UL 94 Vertical Burning Test.

In this test a rectangular bar is held vertically and clamped at the top. The burning behaviour of the bar and its tendency to form burning drips when exposed to a methane or natural gas flame applied to the bottom of the specimen is noted. Methods of specimen preparation and details of the test method are laid down in Underwriters Laboratory Inc. Standard UL 94. Materials are rated 94 V-0, 94 V-1 or 94 V-2 according to the following characteristics:

Table 2.13 Typical limiting oxygen index (LOI) values for polymers*†

Polymer	*LOI* (%)
Polyacetal	15
Polymethyl methacrylate	17
Polypropylene	17
Polyethylene	17
Polybutylene terephthalate	18
Polystyrene	18
Polyethylene terephthalate (unfilled)	21
Nylon 6	21–34
Nylon 66	21–30
Nylon 11	25–32
PPO	29–35
ABS	29–35
Polycarbonate of bisphenol A	26
Polysulphone	30
Polyethylene terephthalate (30% G.F.)	31–33
Polyimide (Ciba-Geigy P13N)	32
Polyarylate (Solvay Arylef)	34
Polyether sulphone	34–38
Polyether ether ketone	35
Phenol–formaldehyde resin	35
Polyvinyl chloride	23–43
Polyvinylidene fluoride	44
Polyamide imides (Torlon)	42–50
Polyphenylene sulphide	44–53
Polyvinylidene chloride	60
Polytetrafluoroethylene	90

* Reproduced by permission from Brydson, *Plastics Materials*, 5th edition, Butterworth-Heinemann.
† Where only a single figure is given it applies to raw polymer, unless otherwise indicated; where there is a range of figures the result is dependent on the grade of material, which may be dependent on the polymer structure or the presence of additives such as glass fibre and flame retardants. The relevant standards are ASTM D-2863 and BS 2782: Method 141B.

94 V-0 Specimen burns for less than 10 s after application of test flame. The wad of surgical cotton beneath the specimen is not ignited. When the specimen stops burning the flame is replaced for a further 10 s; after removal of this flame glowing combustion should die within 30 s. The total flaming combustion time for 10 flame applications on five specimens should be less than 50 s.

94 V-1 Specimen burns for less than 30 s after either test flame application: cotton not ignited. Glowing combustion dies within 60 s of second flame removal. Total flaming combustion time for 10 flame applications on five specimens should be less than 250 s.

94 V-2 As for 94 V-1, except that there may be some flaming particles which burn briefly but which ignite the cotton.

The UL 94 ratings of some selected engineering thermoplastics are given in *Table 2.14*.

The UL codings apply to a *specific grade* of material. The ratings may be dependent on the sample thickness, conditions being more stringent with thinner samples. Many of the so-called engineering thermoplastics now meet the V-0 rating and there has been an increasing tendency to quote minimum sample thickness for which the material will pass the V-0 test.

Table 2.14 The UL 94 ratings of some selected engineering thermoplastics*

Polymer	*Tradename*	*UL 94 rating*
Polycarbonate	Lexan 101	V-2
Nylone 66	Maranyl A100	V-2
Nylon 66 G.F.	Maranyl AD 447	V-0
Polysulphone	Udel P-1720	V-0
Polyethersulphone	Victrex 200P	V-0
Polyethylene terephthalate G.F.	Rynite 530 FR	V-0
PPO	Noryl HS 1000	V-0
	Noryl GFN-2	HB
Polyacetal	Delrin 500	HB
Polyphenylene sulphide G.F.	Ryton	V-0
Polyarylate	Ardel D-100	V-0
Polyether–imide	Ultem	V-0
Polyether ether ketone	Victrex PEEK	V-0
ABS Standard Grades		HB
ABS/polycarbonate alloy	Cycovin KMP	V-0

* Reproduced by permission from Brydson, *Plastics Materials*, 5th edition, Butterworth-Heinemann.

Mention should also be made of two other UL ratings; the 94HB rating for which the sample passes a less severe burn test than the vertical test and the 94 5-V rating which is more severe than the V-0 rating and involves a test in which the sample is subjected to a 5 in. flame five times at 5-s intervals with each exposure lasting 5 s.

2.4.4 Toughness

An important prerequisite of plastics materials in most applications is that they should have an acceptable level of resistance to sudden impact. It is quite common to classify plastics into two types: brittle plastics and tough plastics.

A *brittle* plastic material is one that can only withstand a very small strain before fracture (e.g. 1–2%) so that the area under a stress–strain curve is small. Such materials break cleanly with little permanent distortion of the broken surfaces and such parts may sometimes be repaired by welding or by adhesives. Just because a material is classified as brittle does not mean that it is unsatisfactory. This will depend on the expected operational conditions.

Tough plastics extend by several per cent before fracture. In this case the area under the stress–strain curve is quite large, this being a measure of the energy required to break. In some cases a product may yield irreversibly before or without fracture, rendering it unsatisfactory for its purpose so that a simple consideration of toughness is not enough.

For general comparative purposes toughness is measured by standard impact tests, in particular the Izod impact strength. In this test half of a notched sample is supported and then struck at a fixed point above the notch, on the same side as the notch, by a swinging pendulum. The energy required to break is obtained by considering the amount the pendulum follows through after striking the sample relative to the height of the pendulum before its release. Results may be expressed in a number of ways, but it is to be noted that the impact strength will depend on the cross-sectional area of the sample and also on the angle of the notch. Some materials are very sensitive to the notch-tip radius (*Figure 2.14*).

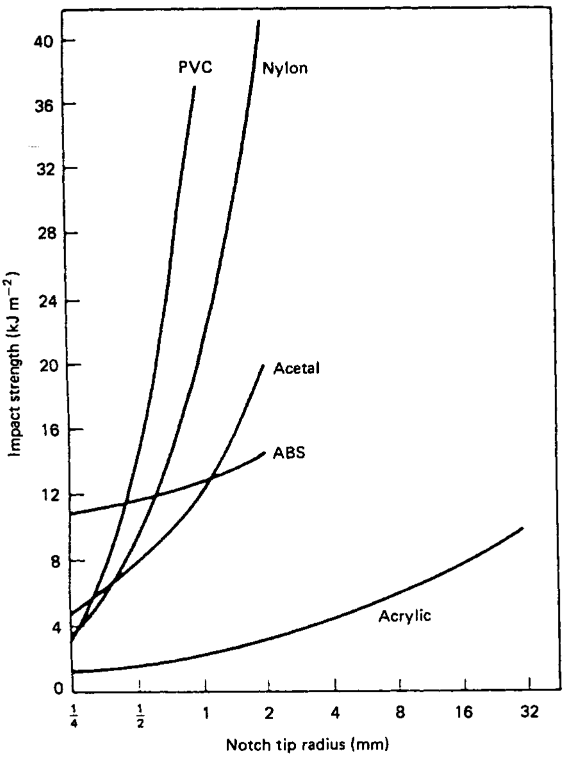


Figure 2.14 Impact strength as a function of notch-tip radius for samples of five different polymers. (Reproduced by permission from Vincent, *Impact Tests and Service Performance of Plastics*, Plastics and Rubber Institute)

Considerable interest has been given to the application of fracture mechanics to plastics since this can help to predict the service life of articles containing small cracks or crack-like defects. For the comparatively simple case of a crack subject to opening forces in a linear brittle elastic material the stress at the crack tip in the situation illustrated in *Figure 2.15* is given by

$$\sigma_Y = K_I/(2\pi r)^{1/2}$$

where r is the distance from the crack tip along the direction of crack growth and

$$K_I = Y\sigma(\pi a)^{1/2}$$

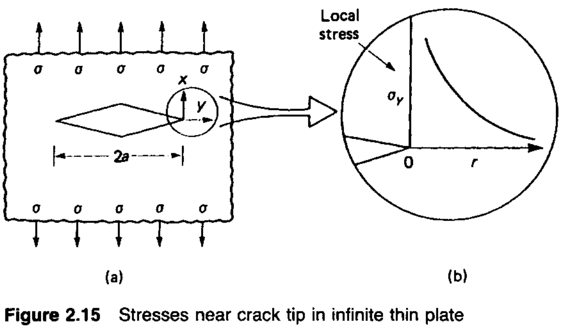


Figure 2.15 Stresses near crack tip in infinite thin plate

i.e. the local stress is proportional to the applied stress and varies with the square root of the crack size. Y is a modifying factor depending on the location of the crack relative to the plate geometry and is obtainable from standard texts. K_I is independent of the material, provided that it is linear elastic. A material will fail by very rapid crack growth when K_I reaches a critical value for the material; typical values of K_{IC} are given in *Table 2.15*.

The crack growth rate is given by

$$da/dt = CK_I^n$$

for materials under constant stress where C and n are constants. These 'constants' may vary with the stage of crack growth. The time for a crack to grow may be found by integration of the crack growth equation between the original size of the crack and the new size of the crack.

2.4.5 Creep behaviour

Under long-term load plastics materials will continue to deform or *creep*. The extent of creep will depend considerably on the type of polymer and also on whether or not it has been reinforced, e.g. with glass fibre. Creep will also depend on the temperature (*Figure 2.16*) and on the load (*Figure 2.17(a)*). A consequence of this creep is that, after an interval of time, the amount of deformation may reach unacceptable levels and this has to be built into design calculations (considered in Section 2.8.1).

It is possible to derive from a family of creep curves such as those shown in *Figure 2.17(a)* two other types of curve. In the first case the strain corresponding to various levels of stress at some specified time, e.g. 1000 h, is noted. The resultant plot is known as an isochronous stress–strain curve (*Figure 2.17(b)*). Any number of such curves may be so derived, simply by changing the specified time.

Alternatively, the amount of stress that will give a specified strain at various times will yield an isometric stress vs. log(time) curve (*Figure 2.17(c)*). Finally, a creep modulus vs. time curve may also be derived (*Figure 2.17(d)*). Although no additional information is actually provided by the derived curves they are sometimes more convenient to use.

It is interesting to note that acceptable upper strain limits are frequently used for plastics materials. Whilst there is some level of arbitrariness, some typical figures are suggested in *Table 2.16*.

On release of a load recovery of plastics material is time dependent and also incomplete, since some of the creep will be irreversible. Data are sometimes provided in *reduced* or *dimensionless form* using the following:

Table 2.15 Typical values for fracture toughness in air at 20°C

Material	K_{IC} (MN m$^{-3/2}$)
Epoxy resin	0.6
Polystyrene	1.0
Cast acrylic sheet	1.6
Polycarbonate	2.2
PVC pipe compound	2.3
High density polyethylene	3.0
Nylon 66	3.6

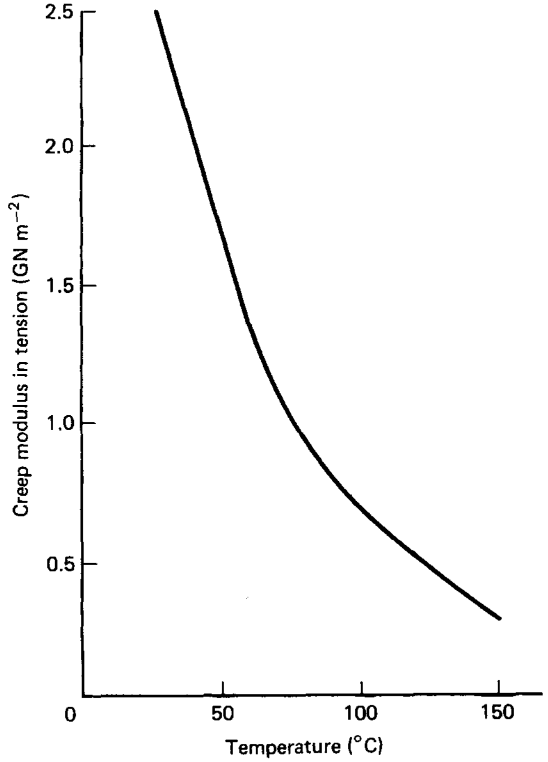


Figure 2.16 Example of the dependence of the creep modulus on the temperature for a crystalline thermoplastic: data typical for a polyacetal copolymer after 100 s loading and 0.2% strain

Table 2.16 Upper strain limits for plastics

Plastic	*Strain* (%)
Glass-reinforced nylon	1
Acetal copolymer	2
Polypropylene	3
ABS	1.5
Polyethersulphone	2.5
Polycarbonate	0.6

$$\text{Fractional recovered strain} = \frac{\text{Strain recovered}}{\text{Total creep strain at moment load is removed}}$$

$$\text{Reduced time} = \frac{\text{Recovery time}}{\text{Duration of creep period}}$$

Typical data are given in *Figure 2.18*.

2.4.6 Solvent resistance

Amorphous polymers tend to dissolve in solvents which have a solubility parameter within 2 $MPa^{1/2}$ of that of the polymer. In the case of crystalline polymers the situation is more complicated and two main effects may be discerned.

If the polymer is not capable of specific interaction with a solvent (as with PE, PP and PTFE) then there are no solvents until temperatures approaching the melting point are reached.

If the polymer is capable of a specific interaction with a solvent then it may be possible to find a solvent if it shows complementary specific interaction. Such interaction is possible with polycarbonates, PVC and the nylons.

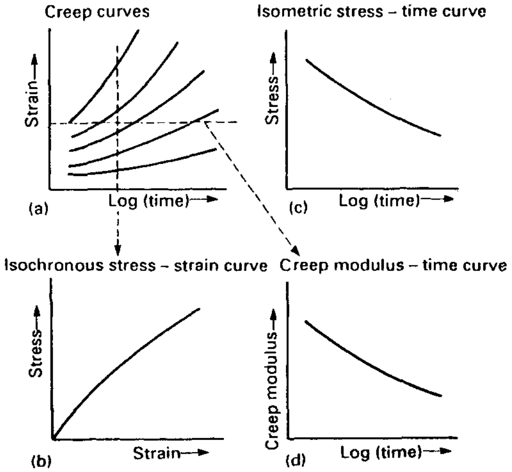


Figure 2.17 Presentation of creep data: sections through the creep cruves at constant time and constant strain give curves of isochronous stress–strain (b), isometric stress–log(time) (c) and creep modulus—log(time) (d). (Reproduced by permission of ICI)

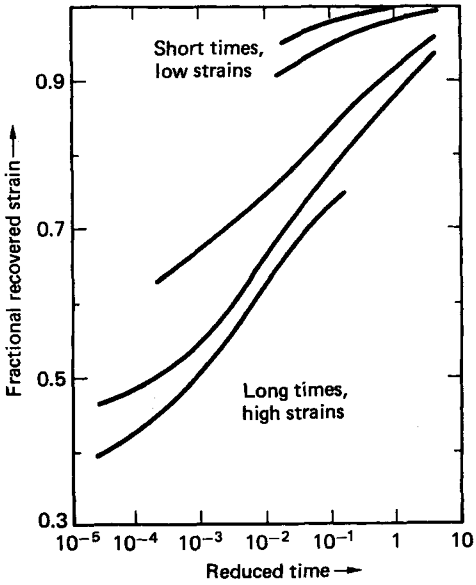


Figure 2.18 Recovery from tensile creep of an acetal copolymer at 20°C and 65% relative humidity. (Reproduced by permission from *ICI Technical Service Note G117*, ICI)

Some solubility parameter data for polymers and solvents are given in *Tables 2.17* and *2.18*. Crystalline polymers marked with a ‡ are not capable of specific interaction and do not dissolve at room temperature, whilst those marked with a § are capable of specific interaction with some solvents and, where these solvents have similar solubility parameters to the polymers, dissolution may occur. In most of these cases it is necessary that one of the components (i.e. polymer and solvent) be what is known as a 'proton donor' and the other a 'proton acceptor'. Examples of proton donors are PVC, dichloromethane, formic acid, cresol and dimethyl formamide, whilst examples of proton acceptors are the nylons, polyethylene terephthalate and other polyesters, the polycarbonates and cyclohexanone.

Cross-linked polymers including vulcanised rubbers will not dissolve but will swell in solvents of similar solubility parameter. The extent of swelling also depends on the level of cross-linking.

Table 2.17 Solubility parameters of polymers*†

Polymer	*Solubility parameter, δ*	
	$(\text{cal cm}^{-3})^{1/2}$	$(\text{MPa}^{1/2})$
Polytetrafluoroethylene‡	6.2	12.6
Polychlorotrifluoroethylene	7.2	14.7
Polydimethyl siloxane	7.3	14.9
Ethylene–propylene rubber	7.9	16.1
Polyisobutylene	7.9	16.1
Polyethylene‡	8.0	16.3
Polypropylene‡	8.0	16.3
Polyisoprene (natural rubber)	8.1	16.5
Polybutadiene	8.4	17.1
Styrene–butadiene rubber	8.4	17.1
Poly-t-butyl methacrylate	8.3	16.9
Poly-n-hexyl methacrylate	8.6	17.6
Poly-n-butyl methacrylate	8.7	17.8
Polyethyl methacrylate	9.0	18.3
Polymethylphenyl siloxane	9.0	18.3
Polysulphide rubber	9.0–9.4	18.3–19.2
Polystyrene	9.2	18.7
Polychloroprene rubber	9.2–9.4	18.7–19.2
Polymethyl methacrylate	9.2	18.7
Polyvinyl chloride§	9.5	19.4
bis-Phenol A polycarbonate§	9.5	19.4
Polyvinylidene chloride‡	9.8–12.2	20.0–25.0
Ethyl cellulose	8.5–10.3	17.3–21.0
Cellulose dinitrate	10.55	21.6
Polyethylene terephthalate§	10.7	21.8
Acetal resins§	11.1	22.6
Cellulose diacetate	11.35	23.2
Nylon 66§	13.6	27.8
Polymethyl-α-cyanoacrylate	14.1	28.7
Polyacrylonitrile	14.1	28.7

* Reproduced by permission from Brydson, *Plastics Materials*, 5th edition, Butterworth-Heinemann.
† Because of the difficulties in their measurement the published figures for polymer solubility parameters range by ±3% from the average figure quoted.
‡ See text.
§ See text.

Table 2.18 Solubility parameters of common solvents

Solvent	*Solubility parameter, δ*	
	$(\text{cal cm}^{-3})^{1/2}$	$(\text{MPa}^{1/2})$
n-Pentane	6.3	12.8
Isobutylene	6.7	13.7
n-Hexane	7.3	14.9
Diethyl ether	7.4	15.1
n-Octane	7.6	15.5
Methyl cyclohexane	7.8	15.9
Ethyl isobutyrate	7.9	16.1
Di-isopropyl ketone	8.0	16.3
Methyl amyl acetate	8.0	16.3
Turpentine	8.1	16.5
Cyclohexane	8.2	16.7
2,2-Dichloropropane	8.2	16.7
sec-Amyl acetate	8.3	16.9
Dipentene	8.5	17.3
Amyl acetate	8.5	17.3
Methyl-n-butyl ketone	8.6	17.6
Pine oil	8.6	17.6
Carbon tetrachloride	8.6	17.6
Methyl-n-propyl ketone	8.7	17.8
Piperidine	8.7	17.8
Xylene	8.8	18.0
Dimethyl ether	8.8	18.0
Toluene	8.9	18.2
Butyl cellosolve	8.9	18.2
1,2-Dichloropropane	9.0	18.3
Mesityl oxide	9.0	18.3
Isophorone	9.1	18.6
Ethyl acetate	9.1	18.6
Benzene	9.2	18.7
Diacetone alcohol	9.2	18.7
Chloroform	9.3	19.0
Trichloroethylene	9.3	19.0
Tetrachloroethylene	9.4	19.2
Tetralin	9.5	19.4
Carbitol	9.6	19.6
Methyl chloride	9.7	19.8
Dichloromethane	9.7	19.8
Dichloroethane	9.8	20.0
Cyclohexanone	9.9	20.2
Cellosolve	9.9	20.2
Dioxane	9.9	20.2
Carbon disulphide	10.0	20.4
Acetone	10.0	20.4
n-Octanol	10.3	21.0
Butyronitrile	10.5	21.4
n-Hexanol	10.7	21.8
sec-Butanol	10.8	22.0
Pyridine	10.9	22.3
Nitroethane	11.1	22.6
n-Butanol	11.4	23.2
Cyclohexanol	11.4	23.2
Isopropanol	11.5	23.4
n-Propanol	11.9	24.2
Dimethyl formamide	12.1	24.7
Hydrogen cyanide	12.1	24.7
Acetic acid	12.6	25.7
Ethanol	12.7	26.0
Cresol	13.3	27.1
Formic acid	13.5	27.6
Methanol	14.5	29.6
Phenol	14.5	29.6
Glycerol	16.5	33.6
Water	23.4	47.7

* Reproduced by permission from Brydson, *Plastics Materials*, 5th edition, Butterworth-Heinemann.

3 Metal Casting and Moulding Processes

J Glownia

Contents

3.1 Economics of casting and moulding 3/3

3.2 Sand casting 3/4
- 3.2.1 Pattern making 3/4
- 3.2.2 Moulding and casting processes 3/5
- 3.2.3 Making a mould 3/6
- 3.2.4 Gating systems 3/9
- 3.2.5 Risers 3/12

3.3 Pressure die casting 3/14
- 3.3.1 Low pressure die casting 3/14
- 3.3.2 High pressure die casting 3/15

3.4 Investment casting 3/17
- 3.4.1 The Shaw process 3/17

3.5 Shell moulding 3/19

3.6 Sintering 3/21

3.1 Economics of casting and moulding

Casting may be defined as 'metal shaping obtained by the solidification of molten alloy in the mould cavity of a required shape'. Of course, there are also other methods of shaping metal, i.e. forging, stamping, rolling, machining or welding. In practice, wrought iron is cast as ingots, and then plastically worked into the required shape. This means that all primary alloy metals are cast first.

The process of casting has certain advantages over the other processes of shaping:

(1) it can be adapted to the requirements of mass production;
(2) castings have more uniform properties from the directional point of view than does wrought iron;
(3) heavy pieces (100–200 ton) can be cast, while it would be difficult to make these in any other way;
(4) the design of parts can be simplified so that machining or forging is reduced to a minimum;
(5) some metals or alloys cannot be hot worked from ingots into other shapes.

The predominant factor in every process is its economic aspect. The design of the product may dictate whether a single process of shaping or a combination of several methods is suitable.

Developments in the production of castings are related to the implementation of new foundry technologies and with achievements in the many branches of industry. A new development requires the use of a material of strictly determined properties that corresponds to the operating conditions that it will experience. Examples of well-defined materials are steels for service at low temperatures, e.g. transcontinental pipelines in Siberia and Alaska, and the new grades of corrosion-resistant cast steel with a carbon content of less than 0.03% which eliminates intercrystalline corrosion. In the latter case interest is focused on materials which are used in aviation and space applications, e.g. composites and monocrystals. Each of the above-mentioned operating conditions requires the use of special-purpose alloys.

Remaining competitive in what is now a world market for cast products requires continuous study and evaluation of the production of the main metals and alloys.

For more than 100 years the production of steel has been systematically growing all over the world (*Figure 3.1*). In 1960, the total production of steel in the world amounted to 341 million ton and in 1987 it increased to 734.7 million ton.

It is interesting to note that in highly industrialised countries such as the USA and the EEC countries, the rate of steel production decreased, while it increased rapidly in countries with new industrial traditions such as Brazil, Australia and South Korea (*Table 3.1*).

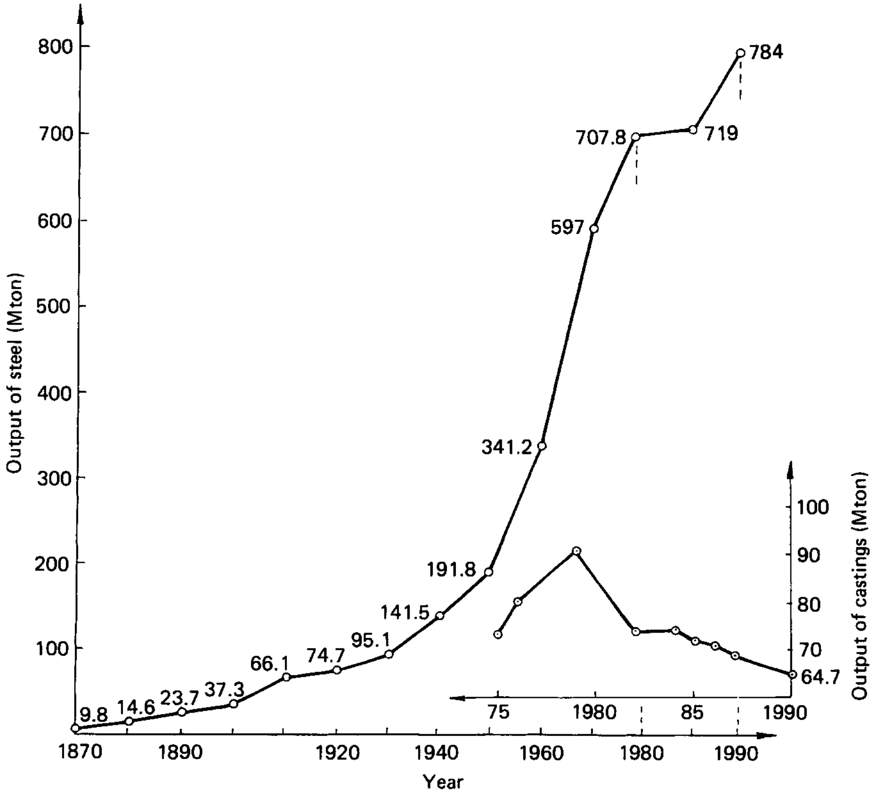


Figure 3.1 World output of steel and castings, 1870–2000

Table 3.1 Production of steel castings (SC) and raw steel (RS) in the years 1970–1987

	1970		*1975*		*1980*		*1982*		*1983*		*1985*		*1987*	
Country	*SC*	*RS*	*SC*	*RS*	*SC*	*RS*	*SC*	*RS*	*SC*	*RS*	*SC*	*RS*	*SC*	*RS*
Australia	—	6900	60	8100	74	7900	75	6400	75	5700	75	6600	—	6100
Austria	26.5	4100	34.5	4500	28.3	5100	—	4300	18.1	4400	25.81	4700	—	4300
Belgium	97.5	42600	88.7	11600	80.7	12300	74.8	10000	76.6	10200	70.5	10700	—	9800
Brazil	—	—	116	—	163.5	15300	109.4	13000	103	14700	133.7	20500	—	22200
Canada	174	11200	217	13000	198.2	15900	117	11900	36.4	12800	80.4	14600	—	14700
Czechoslovakia	300	11500	349	14300	363.2	15200	363.6	15300	342.6	15000	342.6	15000	—	15400
France	288	23800	263	21600	218.2	23700	180.4	18400	142.4	17600	139.6	18800	—	17700
FRG	390	45000	369	40400	293.7	43000	251.8	35900	206.6	35700	228.2	40500	—	36300
GDR	—	5100	204	6500	233.2	7300	237.4	7200	—	7200	228	7900	—	8200
Great Britain	263	28300	284.5	20200	174.1	11500	147.8	13700	114.9	13000	118.7	15700	—	17200
India	—	6300	68	7800	74	9500	74	11000	—	10200	—	11900	—	12600
Italy	154	17200	168	21800	132.9	26500	115.3	24000	84.5	21800	93.8	23900	—	22800
Japan	897	93300	644	102000	732.6	111000	612.9	99500	518.8	97200	544	105300	—	98500
Poland	286	11795	342	15000	354.4	19500	355	14800	266	16200	258	16100	—	17000
Spain	127	7400	142	11100	140	13000	105	13200	99.5	13000	105	14200	—	11900
Sweden	38	5500	37.4	5200	41	4200	12	3900	13.5	4200	14.7	4800	—	4600
USSR	—	116000	5582	141000	323	140000	—	147200	—	152500	—	154700	—	161400
Republic of Korea	—	—	—	—	—	—	92	11800	—	11900	105	13500	—	16800
China	—	—	—	—	—	—	683	37200	—	40000	1000	46700	—	55300
South Africa	—	—	—	—	—	—	141.7	8300	—	720	1077	8500	—	8800
Taiwan	—	—	—	—	—	—	23	4200	—	5000	34.7	5100	—	5600

In recent decades some important changes have occurred in the structural system of steel manufacture. The Bessemer process, discovered in Germany in 1856, and the Thomas process dominated the manufacture of steel in the last century, but are now very little used. The Siemens–Martin process discovered in France in 1865, and the extension to the basic lining in 1875 have also become less important. The steel-manufacturing processes used today are aimed at achieving a higher melting rate with a simultaneous reduction in cost and improved quality of the steel obtained. Currently, attention is focused on the oxygen converter processes. A small but constant growth in popularity has been gained by the electric arc process. The share of these processes in the general production of steel is shown in *Figure 3.2*. Different production processes are used in different countries, depending on the availability of raw materials and the level of technical knowledge.

Future forecasts are based on the use of three processes: Siemens–Martin, electric arc, and oxygen converter. It is anticipated that in the year 2000 the share of the oxygen-converter processes will account for about 70% of the steel manufactured, the electric-arc process for 20%, and the Siemens–Martin process for the remaining 10%. It is expected that Siemens–Martin furnaces will be used only for the manufacture of large castings (50–100 ton).

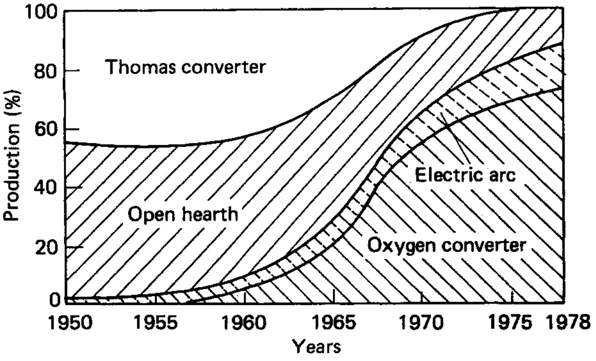


Figure 3.2 Production of steel by various processes. (Reproduced by permission from Graf[15])

Figure 3.3 shows the trends in the amount of various alloys used world-wide. In the last decade there has been a decrease in the production of grey cast iron and cast steel, but an increase in the share of ductile iron and aluminium alloy casting.

3.2 Sand casting

Making castings in sand moulds is a multistage process consisting of the following operations:

(1) pattern making—including core boxes, preparation of the equipment such as drags, flasks and patterns of the gating system;
(2) core and mould making;
(3) melting and pouring of the alloy;
(4) solidification; and
(5) cleaning and inspection.

3.2.1 Pattern making

The main objective of using a pattern is to reproduce the external shapes of a casting. The core reproduces the internal shape of a casting. The pattern equipment also includes a set of gate or riser patterns and patterns of the core boxes in which cores are made.

Three main groups of patterns are used:

(1) *Loose patterns* These are unsplit or split patterns with or without loose pieces. When making patterns allowance must be made for the contraction of the alloy during solidification. Patterns are mainly made of either wood, plastics or alloys such as aluminium and cast iron. They are used in the manufacture of a small number of castings.
(2) *Loose patterns with core prints* The loose part remains in the mould when the pattern is lifted from the mould, and is taken out at a later stage. This enables the pattern to be lifted from the mould without damage.
(3) *Gated patterns* Compared with the above two types, gated patterns are designed in such a way that, in

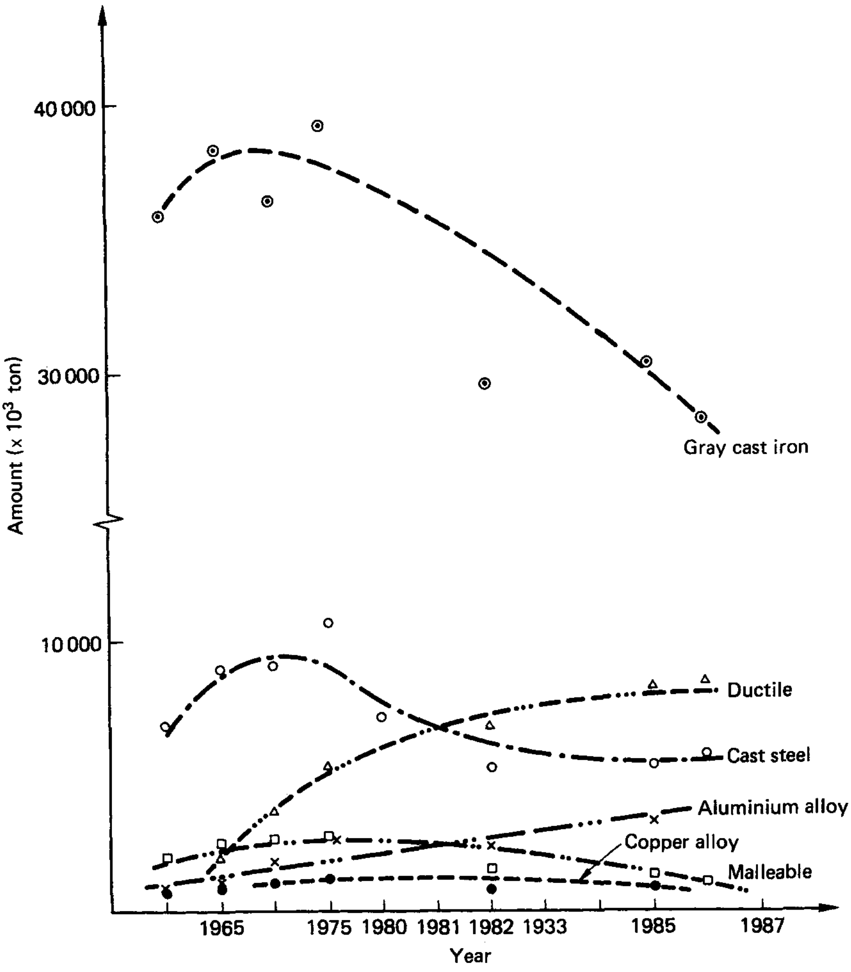


Figure 3.3 Trends in the amount of various alloys used world-wide

addition to reproducing the proper shape of a casting, they also reproduce the gating system and core prints. They are usually made of aluminium alloy or cast iron. They are used to reproduce small castings made in large batches. They are usually fixed to the cope and drag pattern plates. The cope and drag half moulds are prepared separately. This method is more expensive than those described above. However, it is necessary to use patterns of this type in a mechanised moulding shop.

Although a pattern reproduces the shape of a casting, its dimensions vary from those of the casting itself. The differences arise due to contraction during the solidification process and the allowance made for machining the cast surface.

In view of this contraction in the volume of an alloy in the liquid, solid–liquid or solid state, the pattern has larger dimensions than the casting. The amount of contraction depends on the type of alloy cast, the degree of intricacy of the casting shape, the range of cooling temperatures used, and the rate of cooling. Mathematically, the relative contraction (ε) of an alloy is given by

$$\varepsilon = \alpha(T_1 - T_2) \tag{3.1}$$

where α is the coefficient of thermal expansion and T_1 and T_2 are the temperatures (in kelvin) from and to which the casting is cooled, respectively.

The percentage linear casting contraction of an alloy is defined as

$$\varepsilon_f = \frac{l_p - l_c}{l_c} \times 100\% \tag{3.2}$$

where l_p is the length of the pattern and l_c is the length of the casting.

Values of the casting contraction for various cast alloys are given in *Table 3.2*.

3.2.2 Moulding and casting processes

The process of making a casting consists of the following steps:

(1) preparation of charge and moulding materials;
(2) preparation of pattern equipment such as patterns and core boxes;
(3) melting;

Table 3.2 Casting contraction in the main cast alloys

Alloy	*Contraction* Restrained	*Free*	T_L* (K)	T_p† (K)
Grey cast iron			1403	
Small castings	0.9	1.0		1573–1623
Medium-size castings	0.8	0.9		1523–1573
Heavy castings	0.7	0.8		1493–1523
Cast steel			1783	
Carbon cast steel	1.3–1.7	1.6–2.0		1693–1743
High-alloy chromium	1.0–1.4	1.3–1.7		1713–1753
Austenitic	1.7–2.0	2.0–2.5		1613–1653
Ferritic–austenitic	1.5–1.9	1.8–2.2		1713–1743
Malleable cast iron			1773–1848	
Whiteheart	0.8–1.5	1.5–2.5		1653–1723
Blackheart		<1.2		
Brass			1188	
Zinc	1.5–1.7	1.8–2.0		
Silicon	1.6–1.7	1.7–1.8		1323–1423
Manganese	1.8–2.0	2–2.3		
Bronze				
Tin	1.2	1.2–1.8	1253	1273–1423
Silicon				1323–1423
Aluminium				
Silumin	0.8–1.0	1–1.2	843	973–1003
Aluminium–copper	1.4	1.6	903	

* Temperature of melting.
† Temperature of pouring.

(4) proper preparation of the foundry mould and casting process, i.e. pouring of mould and solidification;
(5) cooling of the casting; and
(6) cleaning and final treatment which includes heat treatment and control of the surface and of mechanical properties, cutting of risers and machining.

The term 'mould making' refers to a method of making moulds and to the materials used in the process.

The following guidelines are recommended in the design of parts to be produced by casting:

(1) large surfaces of a casting which are to be machined should be located in a horizontal position in the drag;
(2) the position of a casting in the mould should be such as to ensure easy placing of the cores, and hanging cores should be avoided if possible;
(3) the size of moulding boxes and the position of a casting in the mould should ensure the lowest possible consumption of the sand, hence reducing labour cost;
(4) castings made from alloys which have a high shrinkage value should be moulded in a position that ensures their directional solidification;
(5) the pattern should have one parting plane and no loose parts;
(6) the machining datum should be a rough surface or a surface with the smallest possible machining allowance;
(7) the core box should have sufficent strength to counteract any changes in its position on pouring and allow for the vents and the stability of its shape during mould assembly; and
(8) the placing of several cores in one core print seat should be avoided.

3.2.3 Making a mould

In a traditional process the mould is made by pouring the sand and ramming it around the pattern. Packing of the sand can also be done by hand, a pneumatic rammer, a slinger, jolt squeezing, or squeezing and shooting.

There are three types of mould, the type used depending on the life required:

(1) expendable, single-use type moulds;
(2) semi-permanent moulds; and
(3) permanent moulds, or dies.

The life of a mould is expressed as the number of pourings for which the mould can be used before its destruction.

Expendable moulds are made of natural or synethetic sands. The basic material is silica sand, mixed with a binder which is bentonite, kaolinite clay with an addition of oils, dextrine, waste sulphite liquor, or a resin. *Green sands* are sands with a moisture content of up to 6% and a compression strength of 2–8 × 10^4 Pa, and *skin-dried and dried sands* have compression strengths above 1 × 10^6 Pa.

Semi-permanent moulds are made of highly resistant materials such as *chamotte*, graphite, or plaster. Their life corresponds to several dozen pourings, and for every pouring only the surface layer of the mould cavity is renewed. These moulds are used only for large and medium-sized castings of simple shape that are manufactured in batches.

Permanent moulds are made of cast iron, alloyed cast steel or copper. Their life ranges from several tens of pourings for steel castings up to several thousand pourings for aluminium alloys. They are used in the mass batch production of small and medium-sized castings.

In foundry practice single-use moulds are most popular. Some different types of these moulds are shown in *Figure 3.4*. These types of mould can be made in a box or on the foundry floor. The selection of the type of mould depends on the number of castings, the dimensional accuracy and the surface finish required.

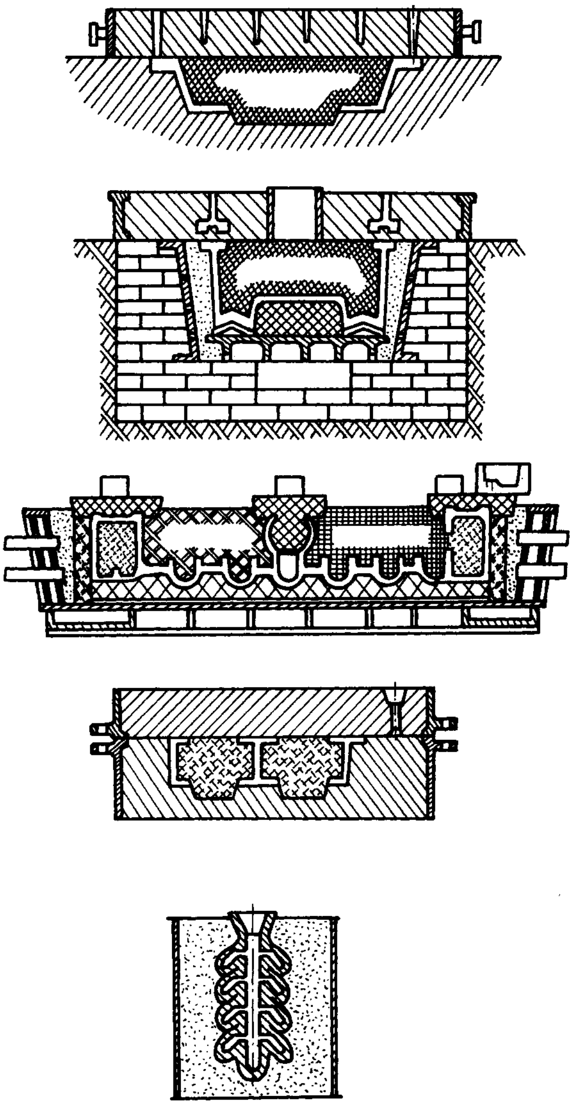


Figure 3.4 Types of single-use mould: (a) foundry floor; (b) caisson; (c) core assembly; (d) boxes; (e) shells. (Reproduced by permission from Skarbiński[8])

When designing single-use moulds the following considerations should be taken into account.

(1) A mould should have sufficient resistance to the impact of the stream of metal and to the metal static pressure; a resistance of this type is ensured through using a sand of appropriate compression and bending strength.
(2) A mould should preserve its refractory abilities from the moment of pouring until knocking out of the casting. The time of solidification for large castings may be 20–40 h and during this time reaction between the molten metal and the moulding sand is avoided by using protective coatings applied to the surface of the mould cavity, or by using more refractory sands for direct contact with molten alloy.
(3) A mould should possess appropriate permeability to enable escape of the gas evolved from the sand during the solidification of an alloy; the permeability is controlled during sand preparation.
(4) A mould should possess sufficient flexibility to control the stresses and shrinkage on solidification.
(5) There should be good collapsibility during the separation of a casting from the mould and crushing of the moulding sand.

The chemical compositions of moulding sands are given in *Table 3.3*.

In modern technologies synthetic sands based on cement, water glass and furan resins are widely used. The main advantage of using such sands is that it is possible to control their properties. Disadvantages include poor collapsibility, difficult reclamation and pollution of the environment.

Self-setting liquid and loose sands, which do not necessitate drying of moulds and cores are currently gaining in popularity. Liquid self-setting sands are poured into a moulding or core box, where they harden without any ramming. Loose self-setting sands, on the other hand, require some packing.

The choice of self-setting sands is based on four criteria:

(1) reducing the time required (and thus the cost) to prepare the mould so that liquid sand can simply be poured into the mould only;
(2) the rate of setting;
(3) ease of knocking out of casting; and
(4) reducing the cost of the sand components used.

A comparison of the relative costs of the manufacture of moulds and cores in various types of sand is given in *Figure 3.5*.

The main operations of the mould-making process consist of the following stages:

(1) placing the pattern and box on the pattern plate;
(2) coating the pattern with a material such as powder, kerosene or paraffin which prevents the sand from sticking to the pattern;
(3) filling the moulding box with a facing and backing sand;
(4) packing the sand;
(5) removing the excess sand and placing the vents;
(6) turning over and lifting the pattern;
(7) operations (1) to (6) for the second half-mould; and
(8) blowing the mould cavity and gating system with compressed air, assembly of the two half-moulds, and clamping.

In practice, all of the above operations can be mechanised. The mechanised method of sand packing consists of shooting and squeezing the sand, as shown in *Figure 3.6*. The advantage

Table 3.3 Composition of moulding and core sands

Type of sand	*Application*	*Composition* (wt%)	*Binder* (wt%)	*Additives* (wt%)
System sand	Machine moulding of castings in aluminium alloys	Sand: 80–95 Silica sand: ≤20	Bentonite: 6–8	—
	Machine moulding of castings in copper alloys	Sand: 70–90 Silica sand: ≤25	Bentonite: 6–8	—
	Machine moulding of castings in grey iron (weight 5–100 kg)	Sand: 85 Silica sand: 10–15	Bentonite: 1.5	Coal dust: 0.6
	Machine moulding of steel castings	Sand: 55–65 Silica sand: 25–30	Bentonite: 7–12	Dextrine: 0.3
Facing sand	Iron castings	Sand: 50–90 Silica sand: ≤38	Bentonite: 7–10	Coal dust: ≤5
	Iron castings (weight ≤2 ton, wall section 30 mm)	Sand: 50–60 Half-strong sand: 30–50	—	Sawdust: 10
	Responsible steel castings of medium weight	Chamotte: 85 vol%	Kaolinite clay: 15 vol%	—
	Hadfield steel castings	Chromite sand: 98–100	Kaolinite clay: 2 vol%	—
	Large castings in aluminium alloy	Sand: 50–60 Half-strong sand: 40–50	—	—
Core sand	Heavy iron castings	Sand: 35–45 Half-strong sand: 35–45	—	Sawdust: 10–15 Coal dust: 5–6
	Steel castings susceptible to cracking	Sand: 20–30 Half-strong sand: 50–60	—	Sawdust: 20 Waste sulphite liquor: 1
	Shell moulds and cores	Fractionated sand: 90–92	Resin: 7–10	Kerosene: 1
	Steel castings, CO_2 hardening	Silica sand: 90–95	Water glass (M = 2.7): 5–10	—
Loose sand	For heavy castings up to 2 ton and wall section above 30 mm	Sand: 20–30 Silica sand: 55–65	Clay: 3–5	Cement: 8–10
	Furan sand	Silica sand: >97	Resin: <2	Orthophosphoric acid
Liquid sand	Self-setting good collapsibility; Synflo sand	Silica sand: <95	Hydrophilic agent: <1 Resin: <3.5	Orthophosphoric acid: 0.5 Sapogen: 0.05

of this method is a high moulding output: up to 150 moulds per hour for horizontal mould joints and 300 moulds per hour for vertical mould joints.

Another way of making the mould stable is the 'V' process, which utilises the physical effect of vacuum (*Figure 3.7*). In essence, this method consists in 'sealing' the pattern with foil of thickness 0.05–0.1 mm and making it fit the pattern and plate through the application of negative pressure using a vacuum pump. A special moulding box with a collector and sucking pumps is put into place and filled with binder-free sand. After a preliminary packing by means of vibration, the second foil is put into place and the air is sucked off from the sand mould. Atmospheric pressure then squeeezes the sand to the required hardness. The negative pressure is maintained

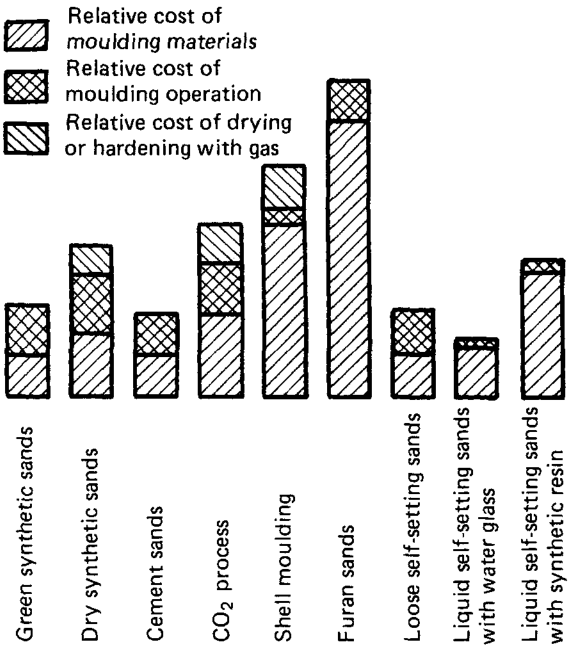


Figure 3.5 Relative cost of manufacturing moulds from various types of sand

until the casting solidifies. When the vacuum is removed the mould spontaneously disintegrates. The main advantage of this process is the total elimination of the use of a moulding-sand mixture. Compared with traditional methods, the vacuum method reduces both the cost of manufacturing a casting and environmental pollution.

3.2.4 Gating system

After the foundry mould has been made, the next stage is to pour the liquid alloy from tilting ladles.

The alloy, previously melted in a furnace, is subjected to a treatment of deoxidation, desulphurising and inoculation, and is then poured into the mould cavity through a gating system. The main objectives of using a gating system are to:

(1) ensure uniform feeding of the molten metal to all parts of the mould cavity;
(2) obtain a proper distribution of temperature over the entire casting during its solidification (simultaneous or directional) and cooling; and
(3) arrest the slag and protect the metal stream from oxidation.

A typical gating system (*Figure 3.8*) is composed of a pouring cup, down-sprue, runner, ingates, risers and flow-off. The place where the metal is fed to the casting is very important for the casting quality. The general rules that should be observed are as follows.

(1) For castings which do not require feeding during solidification (e.g. grey cast iron, bronzes and light alloys), the

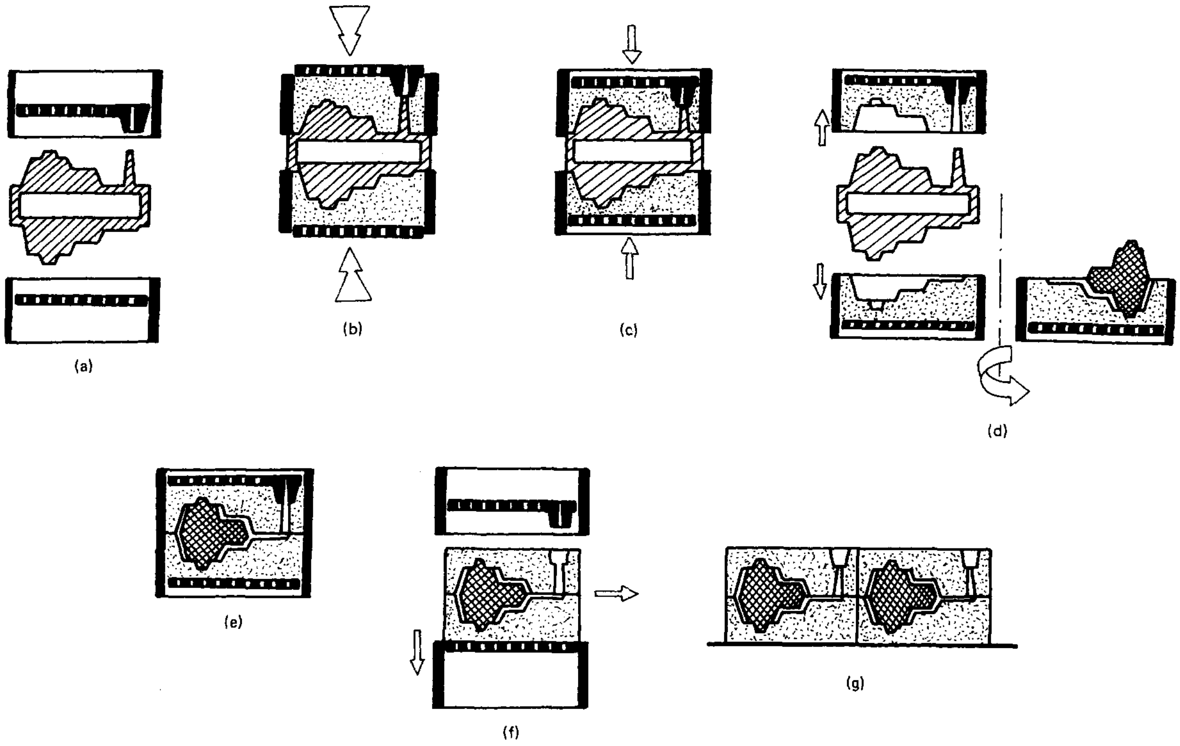


Figure 3.6 Mechanical method of compacting sand in a moulding box. (a, b) Placing of pattern and box; (c) packing of the sand; (d) removal of the pattern; (e) assembly of the two half-moulds; (f, g) before pouring on the moulding line. (Reproduced by permission from Jeancolas *et al.*[6])

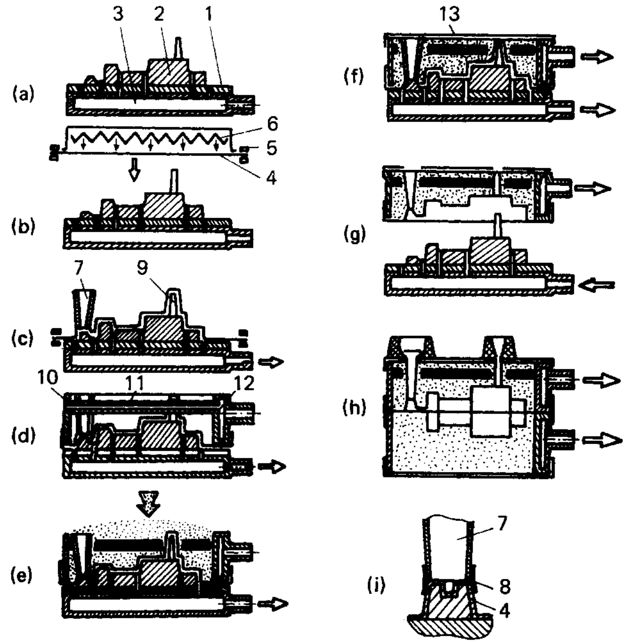


Figure 3.7 Making sand moulds using the vacuum process. (1, 2) Pattern elements; (3) vacuum chamber; (4, 13) foils; (5) foil holder; (6) heating; (7) sprue; (8) sand; (9) riser; (10) box; (11) tubes; (12) collector. (a) Preparation of pattern plate and boxes; (b) heating of the foil; (c) fitting of the foil; (d, e) filling with sand; (f) vacuum turned on; (g, h) assembly of the moulds; (i) moulding of the sprue. (Reproduced by permission from Lewandowski[7])

metal is fed to the thin sections; a great number of the ingates of small cross-section and quick pouring ensure simultaneous solidification.

(2) For heavy castings which require intensive feeding during their solidification (e.g. cast steel, malleable and ductile cast iron, brasses, and copper–aluminium and copper–nickel alloys) the metal must be fed to the thick walls of a casting; thus it is advantageous to feed the metal through feeders and to ensure slow pouring of the mould. This will give directional solidification.

Calculations associated with the gating system consist in determining the pouring time, the pouring temperature, and the sprue–runner–gate ratios. The time of pouring for ferrous alloys is given in *Table 3.4*. The results reported therein were obtained assuming that optimum conditions for filling the mould cavity exist only for a shorter time than the time for alloy cooling in the narrowest cross-section of the mould cavity, in the range of the pouring and liquidus temperatures.[6]

The starting point for the calculations associated with gating systems is to determine the optimum time of pouring (t) or the optimum speed of the metal rising in the mould (V_f). In the case when moulds are poured from tilting ladles there are numerous formulae, both theoretical and practical; for example:

$$t = s \cdot \delta^2 \text{ for cast iron}^1$$

where δ *is the wall thickness,* s *is a coefficient* (s = 12.6 for green-sand moulds; s = 23.0 for dried-sand moulds).[2]

The calculated time of pouring is used to calculate the speed of the metal rising in the mould cavity:

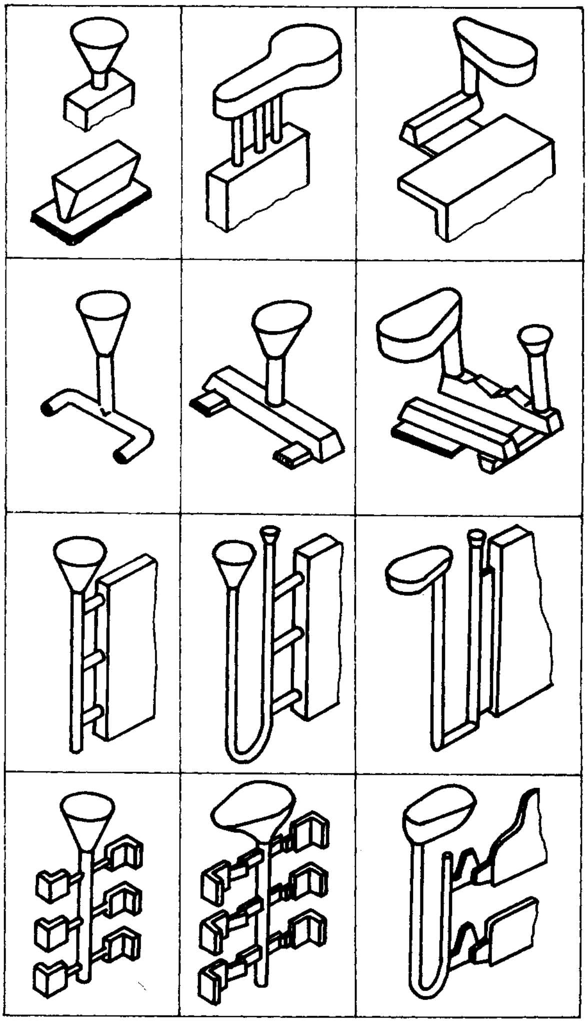


Figure 3.8 A typical gating system. (Reproduced by permission from Skarbiński[5])

$$V_f = \frac{C}{t}$$

where C is the height of the mould cavity.

The calculated value of V_f is checked with the values of V_f given in the literature (*Table 3.4* and *Figure 3.9*).

Once the pouring time has been determined, the smallest cross-section of all the elements of a gating system is calculated using

$$A_i = \frac{Q}{tm} \text{ for cast iron}$$

and

$$A_s = \frac{Q}{tm_1} \text{ for non-ferrous alloys}$$

where A_i is the cross-section of the ingates (cm), A_s is the cross-section of the sprue (cm), Q is the mass of alloy in the

mould (kg), and m and m_1 are mould-pouring-rate coefficients (kg $cm^{-2} s^{-1}$).

The value of A_i or A_s is used as the basis for calculating the remaining dimensions of the gating system. For this purpose the practically tested values of the sprue–runner–gate ratios $A_s : A_r : A_i$ are used. Values of this ratio are given in *Table 3.5*.

Table 3.4 Optimum values of the linear mould pouring rate (cm s^{-1}) for a height of C = 10–150 cm^6

	δ (mm)				
Alloy	*4*	*4–7*	*7–10*	*10–40*	*>40*
Cast iron	3–10	2–3	2–3	1–2	0.8–1
Cast steel	—	—	1	1	0.8
Light alloys	3–10	2–7	1.5–4	0.8–2.8	0.5–1.4

Steel castings are characterised by different conditions of pouring. The use of bottom-stoppered ladles instead of tilting ones is characterised by a low flexibility in the outflow of the liquid steel from the ladle. The discharge of liquid steel is limited by the diameter of the tapping hole in the ladle and by the height of the column of liquid steel in the ladle. Therefore, the size of the ladle and the diameter should be adjusted to the size of the castings.

The formulae used to determine the pouring time are:

$t = a\sqrt{G}$ for castings of weight $G \geq 15$ ton

$t = b^3\sqrt{\delta G}$ for castings of weight $G \leq 15$ ton

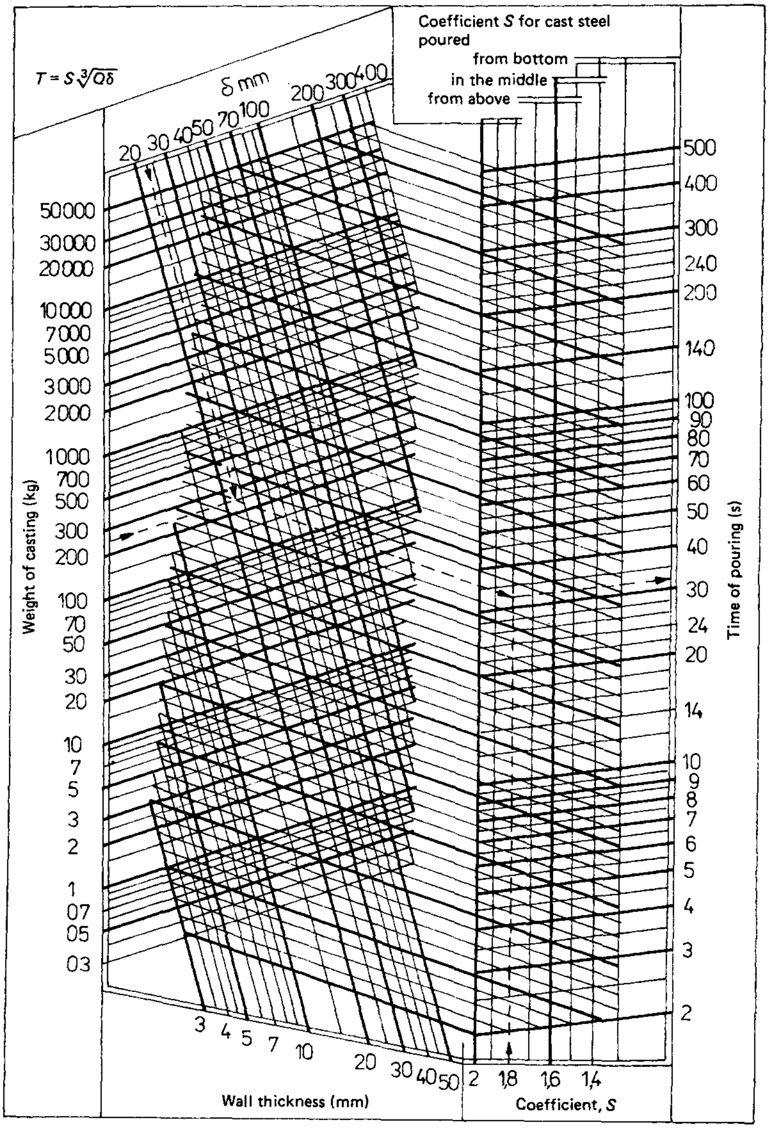


Figure 3.9 Nomogram for calculating the dimensions of a gating system. (Reproduced by permission from Baranowski[14])

Table 3.5 The ratio of the cross-sectional areas of the sprue to that of the runners and ingates

Alloy	A_i	A_r	A_s
Grey cast iron	1	1.1–1.4	1.2–1.6
Malleable cast iron	1	1–1.2	1–1.3
Ductile cast iron	1	1.4	1.5–1.8
Cast steel	1	1–1.2	1.1–1.4
Copper alloy	1	2	1
Copper–aluminium alloy	2–4	2–4	1
Aluminium alloy	2–6	1.2–2	1

where a, b, and p are coefficients and G is the casting weight (kg).

Due to the low flexibility of the system, the down-gate must receive the whole volume of the alloy flowing out from the ladle. Hence the discharge of the liquid steel from the ladle (Q_l) and the volume received by the mould (Q_f) are equal:

$$Q_l = Q_f = \frac{G}{t} = \mu_l\sqrt{2gH_a}A_l\rho = \mu_s\sqrt{2gH_g}A_s\rho \qquad (3.3)$$

where A_l and A_s are the cross-sectional areas of the tapping holes in the ladle and the down-gate, respectively, μ_l and μ_s are coefficients of the flow resistance in the ladle (0.8–0.95) and the mould (0.25–0.5), H_a is the mean metalostatic pressure of the alloy in the ladle, and H_g is the static pressure of the metal in the mould.

After calculating the diameter of the tapping hole in the ladle (25–55 mm) and rearranging equation (3.3), the cross-sectional area of the down-gate is given by

$$A_s = A_l\frac{\mu_l}{\mu_f}\sqrt{\frac{H_a}{H_f}}$$

In the further course of the calculations the sprue–runner–gate ratios (*Table 3.5*) are used.

3.2.5 Risers

During solidification of an alloy, shrinkage cavities are formed in those parts of a casting which solidify at the lowest rate (hot spots). These parts must, therefore, be fed with molten metal. This is possible through proper control of the solidification process in such a way as to locate the shrinkage cavities outside the hot spots in the feeding part. This is done by means of risers connected directly to the hot spots. Risers fulfil their task only when:

(1) there is directional feeding (solidification) from that part of the casting which is most remote from the foot of riser;
(2) the time of action of the riser (time of solidification) is longer than the time of solidification of a hot spot; and
(3) the liquid alloy in the riser absorbs gas and non-metallic inclusions.

In addition, risers must:

(1) be easy to remove and not hinder the transfer of heat from the casting (overheating of the moulding sand); and
(2) not hinder the casting contraction.

The parameter which characterises in a consistent way the process of casting solidification is called the *modulus of solidification* (M). This is determined[3] as the casting volume/cooling surface ratio M_o or as the ratio between the casting cross-sectional area (hot spot) and the circumference of this cross-section, M_j. The solidification modulus is sometimes replaced by the concept of 'reduced wall thickness'.

For intricate castings the calculation of the modulus M appears at first to be somewhat complicated. However, the problem is simplified because:

(1) the value of M for the whole casting is only calculated for large, simple castings; and
(2) in the majority of cases M is calculated only for that part of a casting which acts as a hot spot.

The casting solidification modulus M_o is given by

$$M_o = \frac{V_o^2}{F_c^2} \text{ or } M_j = \frac{F_j}{O_j}$$

where V_o is the casting volume, F_c is the cooling surface of the casting, F_j is the cross-sectional area of the hot spot, and O_j is the circumference of the hot spot.

To preserve directional solidification, the solidification modulus of a riser M_r should have values higher than $M_o \times M_r = 1.2M_o$ to $1.3M_o$ for cast steel, non-ferrous alloys and nodular graphite cast iron, and $M_r = 0.2M_o$ to $0.8M_o$ for grey cast iron, where M_r is the ratio of the riser volume to the cooling surface of the riser.

The calculation of most riser volumes can be done using the reduced formula

$$V_r = M_pF_c + V_{rg} \qquad (3.4)$$

where V_r is the riser volume, M_p is the modulus of the protected part (the riser is divided into two parts—the protected part and the fed part), and V_{rg} is the volume of the fed part of the riser.

Equation (3.4) can be used to calculate the riser volume required for an arbitrary alloy and for an arbitrary shape of riser, providing that the riser cooling rate is low. The calculation of riser dimensions has now been simplified by the availability of nomograms or computer programs,[4] which can be used to calculate and check in practical application the volume of risers. Results are compiled in the form of tables giving the shape and size of risers. An example for risers having a height/diameter (h/d) ratio of 1.5 is shown in *Figure 3.10*.

The next operation is to calculate the number of risers. Although the riser should feed each hot spot which occurs in a casting with metal, for flat and long walls use of a riser does not ensure the elimination of shrinkage cavities since the range of its effect is limited. It is assumed that the range of effect of a single riser is equal to $4M_o$. Section A in *Figure 3.11* is free of any shrinkage defects, due to the boundary effect (heat transfer through a mould) that acts there. The range of effect of the riser depends on the cast wall thickness and on the type

Table 3.6 Maximum feeding distance of risers*

Alloy	*Wall thickness* (mm)	*A*	*K*	*A + K*
Carbon cast steel	<25	2.5d	3.5d	6d
< 0.3%C	50–200	2.5d	2d	4.5d
0.3–0.4%C	15–40	2.5d	3d	5.5d
Malleable cast iron	6–10	—	—	11d
Bronze	10–40	2d	2d	4d
Aluminium alloy	25	—	—	5d

* For definition of A and K see *Figure 3.11*.

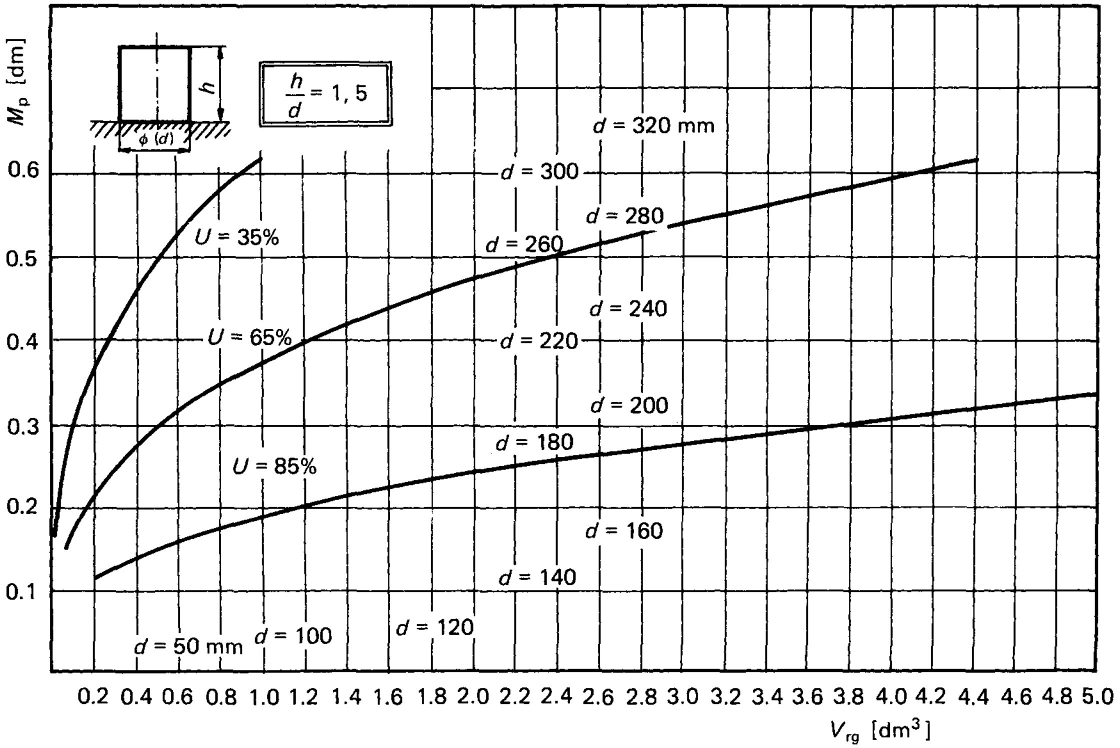


Figure 3.10 Nomogram for calculating the dimensions of a riser. (Reproduced by permission from Wlodawer[4])

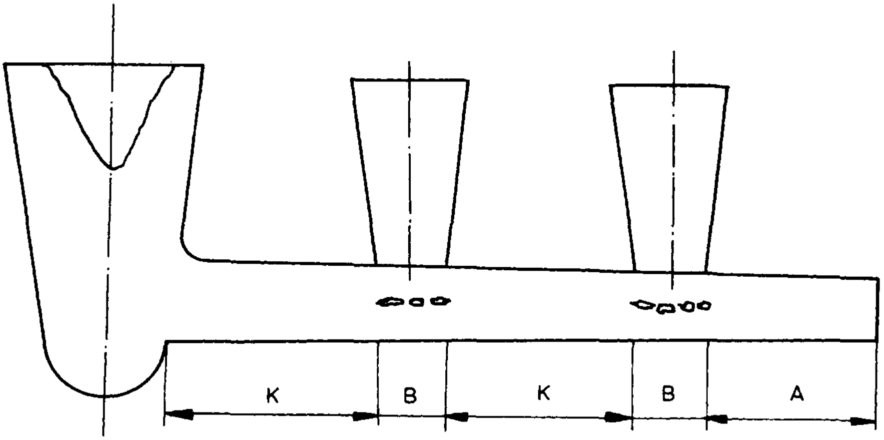


Figure 3.11 The range of effect (*K*) of a riser. (Reproduced by permission from Skarbiński[8])

of alloy; the height of the riser does not play any part in its range of effect (*Table 3.6*).

3.3 Pressure die casting

Pressure die casting is one of the few technologies which in recent years have shown definite growth in usage. This is due to advantageous market conditions in the automotive, electronics and household goods manufacturing industries.

Pressure die casting involves exerting pressure on the liquid metal. Filling of the metallic mould, referred to as a die, is achieved due to a positive gauge pressure (10–70 MPa) of air or nitrogen in the case of low pressure die casting, or by means of a mechanical plunger, which can exert up to 200 MPa, in the case of high pressure die casting.

Pressure-casting dies are most often used to manufacture aluminium alloy, zinc alloy or copper alloy castings; they are also used for casting steel. In the majority of cases dies are made of steel. The die cavity is made of steel characterised by a strength of 700–1100 MPa for low-melting-point zinc alloys or 1200–1600 MPa for copper alloys. These are usually steels containing 1–5%Cr and 0.3–1.1%Mo or 11–12.5%Cr and 0.3–1.2%Mo, in addition to nickel and vanadium.

Compared with other foundry technologies, the pressure die casting process has a number of advantages:

(1) a very high production rate (up to 500 pourings per hour);
(2) good dimensional tolerances (0.03–0.07 mm);
(3) it is possible to cast thin-walled (0.3 mm) details; and
(4) high surface finish and high yield of metal (up to 95%) are obtained.

These characteristics of the process enable elimination of machining and fettling of castings and as much as 40% reduction in the weight of castings. However, the process also has some disadvantages, e.g. high capital costs of the pressure die casting machine and of the new dies, and its application is limited to low-melting-point alloys and castings of low weight.

3.3.1 Low pressure die casting

This process is used for the manufacture of castings which would require the use of large risers in gravity die casting and for thin-walled intricate castings. It is most widely used in the automotive industry (engine cylinders, heads and cylinder blocks) and in the electronics industry (parts of motor rotors).

The advantages of low pressure die casting are:

(1) higher productivity than gravity die casting;
(2) easy mechanisation;
(3) lower investment outlays for machining and fettling; and
(4) a reduced gating system (risers).

There are two types of low pressure die casting machine. In the first type gas pressure is exerted only on the liquid alloy and the walls of the crucible. The machine includes the following elements: a furnace, a timing device, and a gravity die casting machine. Usually, resistance or main frequency induction furnaces are used as the crucible. The timing device is an auxiliary unit which serves to feed compressed air from the cylinder or central supply system, to reduce and stabilise its pressure and, after solidification of the casting, to remove the applied gas. Modern designs are equipped with automatic pressure-controlling devices and appropriate time relays. They control the whole casting cycle and facilitate the operation of machines.

The crucible in a furnace is sealed with a plate on which the casting machine is placed. Under the effect of positive gauge pressure, the liquid alloy is squeezed out of the crucible and fills the die cavity. Since the pressure does not drop, contraction during solidification is compensated for by permanently supplied liquid alloy—this either eliminates or reduces the need for risers.

The metallic mould (die) is made of cast iron and the cores may be made from sand mixture or steel. Since it is possible that cores made of moulding sands might suffer some deformation, the velocity of flow of the alloy in the gate (*Figure 3.12*) should not exceed 1.5 m s^{-1}. The die cavity is coated with a layer of TiO_2 or Al_2O_3 with an addition of $CaCO_3$, while the die is preheated to a temperature of 573–723 K. The temperature used depends on the wall thickness of the casting and on the type of alloy cast.

In the second type of machine, gas pressure is not only exerted immediately above the liquid alloy, but also inside the whole furnace. An example of such a design is shown in *Figure 3.13*. The furnace is equipped with two crucibles. The external cast iron crucible encloses a graphite crucible in which the liquid alloy is held. The temperature is measured at two places: with an immersion thermocouple in the graphite crucible, and in the space between the cast iron crucible and the furnace lining (4 in *Figure 3.13*). The molten alloy flows through a cast-iron pipe (5) heated from the outside by a heating coil. The furnace chamber is sealed by means of a lid (6) and the asbestos lining which adheres to a flange of the cast-iron crucible (2). The cycle of pouring is mechanised and is based on automatic adjustment of alloy temperature and on a control system which includes the time relays.

Automation of the manufacturing cycle usually covers the assembly and removal of a die and filling of the die cavity. The process is controlled through measuring the pressure above the level of molten alloy, the temperature of the molten alloy, and the level of the alloy in the crucible.

Modern machines for low pressure die casting are equipped with a die pouring rate control system. The principle of operation of these machines is by casting with counter-pressure. A characteristic of this method is the possibility of pouring the alloy into a die at a pressure higher than atmospheric pressure. A variation of this process is shown in *Figure 3.14*. At the starting position in this process, the compressed air is fed to an area above the surface of the molten metal (P_1) as well as to the die (3). By turning the device to position II, the die cavity is filled with molten metal due to the effect of gravity.

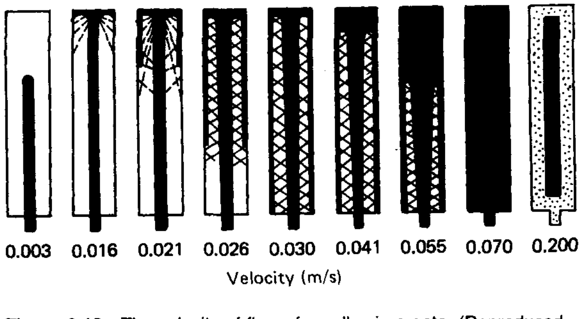


Figure 3.12 The velocity of flow of an alloy in a gate. (Reproduced by permission from Frommer and Lieby[11])

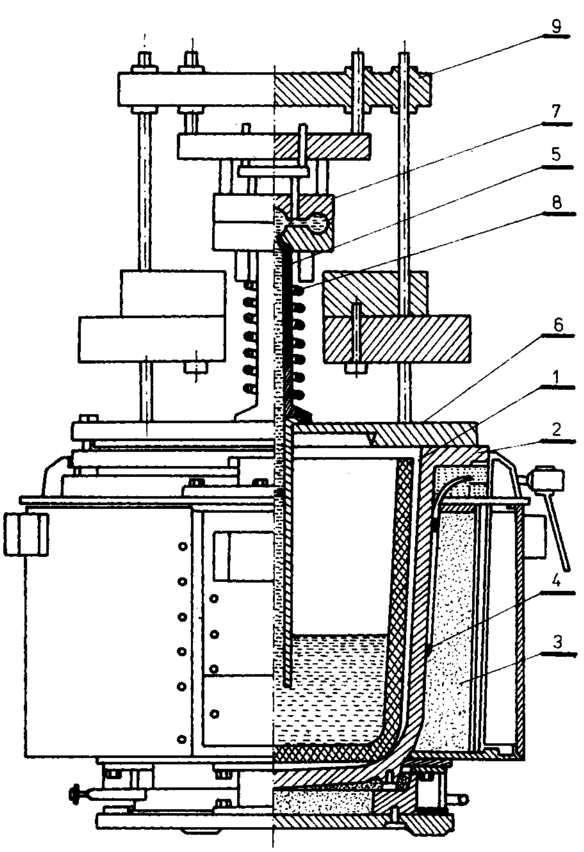


Figure 3.13 Low pressure die casting machine. (1) Crucible; (2) external crucible; (3) housing of the furnace; (4) thermocouple; (5) heat resistant tube; (6) cover; (7) chill; (8) heating; (9) moveable plate. (Reproduced by permission from Green[9])

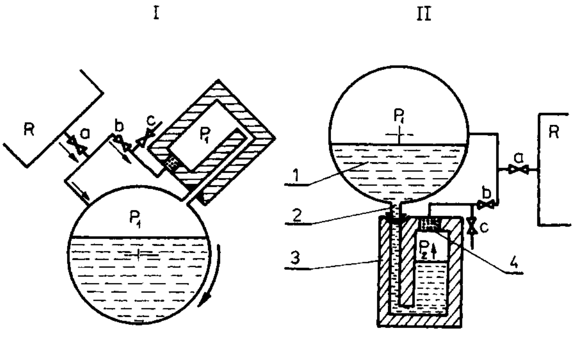


Figure 3.14 Casting of molten metal by counter-pressure. (1) Crucible with molten alloy; (2) bottom outlet; (3) mould. (a–c) Valve system for the compressed air. (P_1, P_2) Pressure ($P_1 > P_2$), (R) pressure vessel. (Reproduced by permission from Balewski and Dimov[10])

3.3.2 High pressure die casting

Unlike in the gravity die casting process, in high pressure die casting machines the liquid metal flows into the die cavity at a pressure of 7–70 MPa. There are two types of machine: hot chamber machines for low-melting-point lead, tin and zinc alloys, and cold chamber machines used for aluminium, magnesium and copper alloys. In the former type of machine it is possible to exert a pressure of 7–15 MPa on the liquid metal, while in cold-chamber machines the pressure can exceed 200 MPa. The design of a die is determined by the temperature of pouring, the alloy castability, the shape and dimensions of a casting, and the type of pressure die casting machine used. Nowadays, the most popular machines are of the hot chamber plunger type in which a high pressure (10–60 MPa) is used.

The plunger is usually hydraulically driven, which makes it possible to exert a pressure of up to 300 MPa on the liquid alloy with the die locking force amounting to 100–5000 ton.

In cold-chamber machines the liquid alloy is kept in a separate furnace outside the machine. An exact quantity of liquid metal is carried in a bucket and poured into the pressure chamber.

In hot-chamber machines the liquid alloy is kept in a crucible directly connected to the pressure chamber. Pressure is exerted on the liquid metal in the crucible, and the metal is then transported under a pressure to the die chamber. In these machines the liquid alloy remains in contact with the plunger and the chamber, which affects the machine operation. This restricts the application of hot-chamber machines, mainly to the production of zinc castings.

The principle of operation of cold-chamber machines is shown in *Figure 3.15*. The liquid alloy is poured into the pressure chamber located in the horizontal position. Under the effect of pressure the plunger pushes the excess alloy out of the chamber in the form of a disk.

Horizontal-chamber machines have the following advantages compared with vertical-chamber machines:

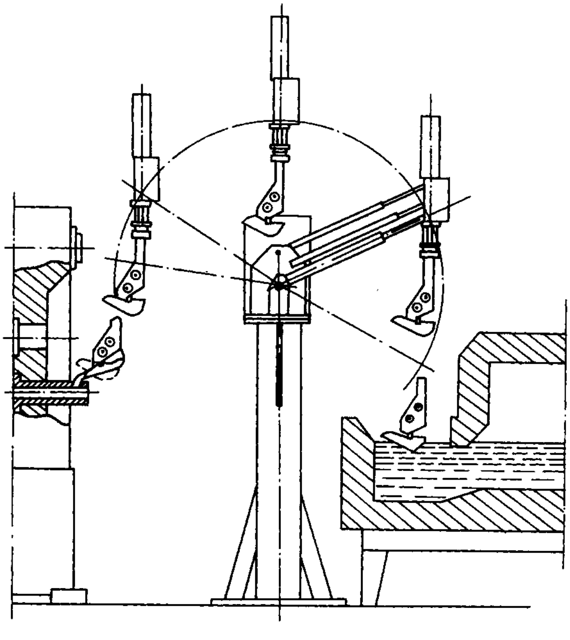


Figure 3.15 Principle of operation of a cold-chamber machine. (Reproduced by permission from Dańko[12])

(1) the molten alloy has a shorter path to pass through from the pressure chamber to the die cavity;
(2) the plunger has better guidance because it does not leave the chamber;
(3) the lubrication of the plunger is easy to mechanise; and
(4) the yield of metal is higher because of the smaller gating systems.

The limitations of horizontal-chamber machines are their greater overall dimensions, and the necessity to provide a better venting system for the die cavity.

A typical pressure die casting machine is shown in *Figure 3.16*. The technological process of making a pressure die casting consists of the following stages:

(1) preparation of the die cavity, including venting and lubrication;
(2) filling the die with liquid alloy; and
(3) consolidation and ejection of the casting.

With regard to casting quality, the consolidation stage is most important. The main task of the squeezing system is to control the increase in pressure during the filling of the die cavity. This factor is very important in view of the way in which the die is filled and the application of pressure.

The following is a typical cycle in the pressure die casting process:

(1) the molten alloy strikes the opposite die wall and its stream is divided into streams flowing in the opposite direction;
(2) the stream of molten alloy strikes the opposite wall and loses its kinetic energy, resulting in a whirling motion of the metal;
(3) the filling of the die proceeds from the opposite wall with intensive whirling of the molten alloy.

Such a die-filling mechanism requires certain operating conditions of the machine and the squeezing system. The most important factors which determine the operation of a pressure die casting machine are:

(1) the locking force and unit squeeze pressure, as they determine the choice of machine power according to the size of the casting;
(2) the ratio between the casting volume and the volume of the liquid alloy introduced into the chamber, as this determines the dimensions of the die and the pressure chamber;
(3) a constant pressure during pouring and solidification.

The three successive stages in an operating cycle of the squeezing mechanism are shown in *Figure 3.17*. The vertical axis in this figure refers to the pressure (P) and the plunger travel (S) and the horizontal axis to time (t).

Stage I: taking the air off from the chamber of consolidation when the pressure exerted on the liquid alloy is still low (1 in *Figure 3.17*).

Stage II: increase in pressure and increase in plunger travel speed (1–3 in *Figure 3.17*). Point 3 denotes the end of filling and the termination of plunger travel.

Stage III: this stage begins at point 3 with the turning on of the hydraulic pressure multiplicator and ends at the moment when the required pressure P is reached (point 4 represents static pressure). This is the consolidation stage (point 5).

An example of a system that controls the operation of the machine is shown in *Figure 3.18*. The system controls the plunger travel speed, the shot time, the consolidation time and the pressure level.

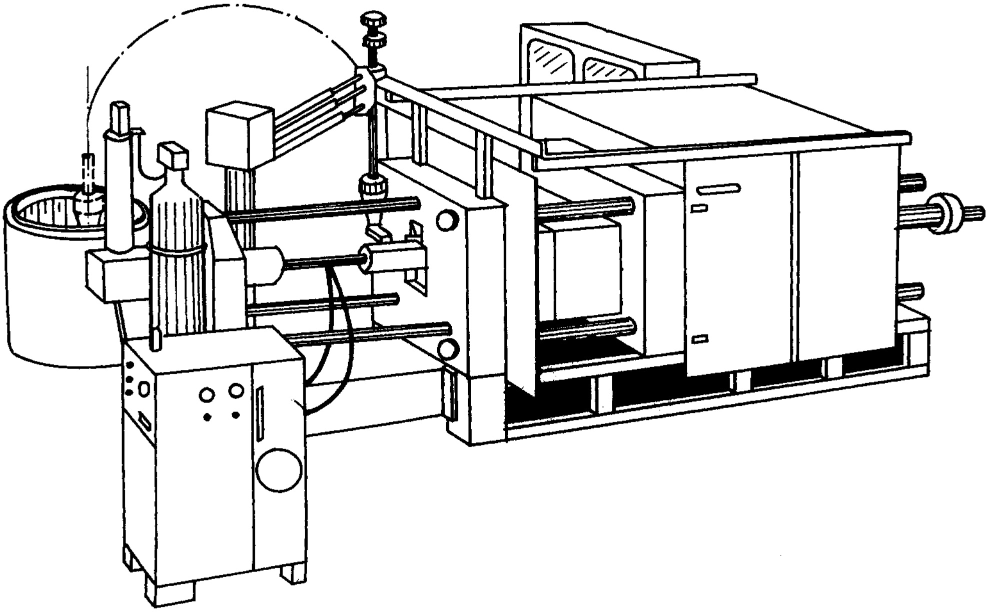


Figure 3.16 A typical pressure die casting machine. (Reproduced by permission from the Bühler Catalog[16])

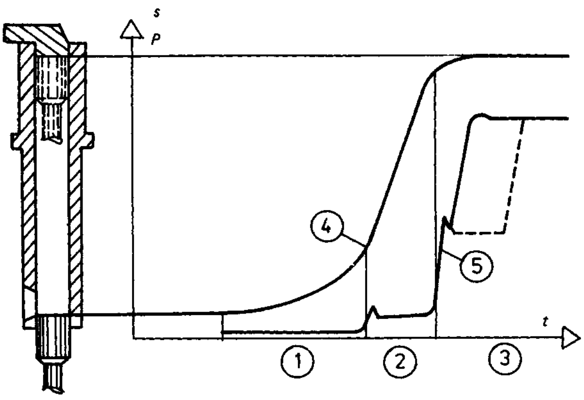


Figure 3.17 A typical operating cycle. For further details see text

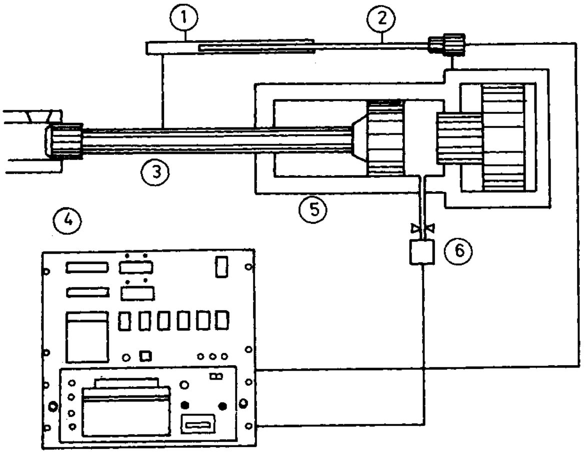


Figure 3.18 Control system for a pressure die casting machine. (1) Cylinder; (2) speed indicator; (3) plunger rod; (4) control unit; (5) shot cylinder; (6) valve

3.4 Investment casting

The investment, or 'lost-wax', process is widely used in the manufacture of intricate castings of low weight. This technology is particularly suitable for batch production. Surface reproducibility is so good that castings need no further machining. With this method it is possible to obtain castings of wall thicknesses as thin as 0.5 mm.

The investment casting process has the following features:

(1) the use of expendable patterns;
(2) moulding in a slurry;
(3) moulds are bonded with an inorganic ceramic binder which ensures exact reproduction of the pattern;
(4) the wax is melted out of the hardened mould;
(5) the mould is heated to remove gases; and
(6) pouring is done into a preheated mould.

The patterns are removed from the wax together with the gating system and risers. The wax patterns with gating systems are made in metallic moulds by injection or pouring. The gating system allows for shrinkage of both the wax and the alloy, and for the expansion of the mould on preheating.

Patterns are made of a mixture of beeswax, paraffin, ceresin or other resins, usually in the form of blends, e.g. 50% paraffin and 50% stearin.

The wax mixture is injected into the chill at 333–343 K. After 3–10 min the pattern is lifted from the chill. When the whole pattern assembly is to be intricate in shape, the patterns and the gating system are made separately. After lifting out the pattern, the flashes are removed. The single patterns of the castings are joined with the gating system to form a cluster (*Figure 3.19*). The patterns and the gating system are joined together by heating and surface melting of the parts which are to be stuck to each other.

The manufacture of a mould consists of the following steps:

(1) preparation of the slurry;
(2) coating of the pattern cluster with a few layers of the ceramic coating;
(3) melting out of the pattern;
(4) moulding of a ceramic shell; and
(5) preheating of the mould.

The slurry, of a refractory coating material, is prepared from ethyl silicate solution, distilled water, acetone (or ethyl alcohol) and hydrochloric acid. Examples of such compositions used for coated-wax assemblies are given in *Table 3.7*.

The pattern assembly is immersed in the slurry several times in order to obtain a uniform coating. After the pattern has been removed from the slurry it is coated by sprinkling with coarse sand and allowed to dry for 4 h in air. During drying the coating hardens due to coagulation of silicic acid and evaporation of acetone and water.

To confer upon a mould certain durability, it is necessary to apply further coatings. The number of coatings applied depends on the weight, shape and degree of complexity of the casting and on the material cast. Usually 3–8 layers are applied. The second and subsequent coatings are of the materials listed in *Table 3.8*.

Successive layers of coating can also be applied by immersing the coated-wax assembly in the slurry on a vibrating table. As the table vibrates, the moulding material surrounds the patterns. Each new layer is sprinkled with sand and then allowed to set in air.

The next operation consists of melting the wax out of the assembly. This is done at a temperature of 353–363 K for 15–30 min. If melting out is done in water, the moulds must subsequently be dried. For drying and preheating, empty moulds are put into containers, surrounded with clean sand and heated to 1150–1320 K (ferrous alloys) or 920 K (aluminium alloys).

The temperature of preheating of moulds is determined by the pouring temperature required for a given alloy and the casting design. The heating and melting out operation must completely remove wax and gas-forming particles from the mould.

Casting is carried in by means of either a gravity feed process or a centrifugal method. After cooling, the casting is cleaned.

3.4.1 The Shaw process

The Shaw process is a variation of the investment casting process. Its features are:

(1) excellent reproduction of the casting shape, and
(2) very good accuracy and surface finish.

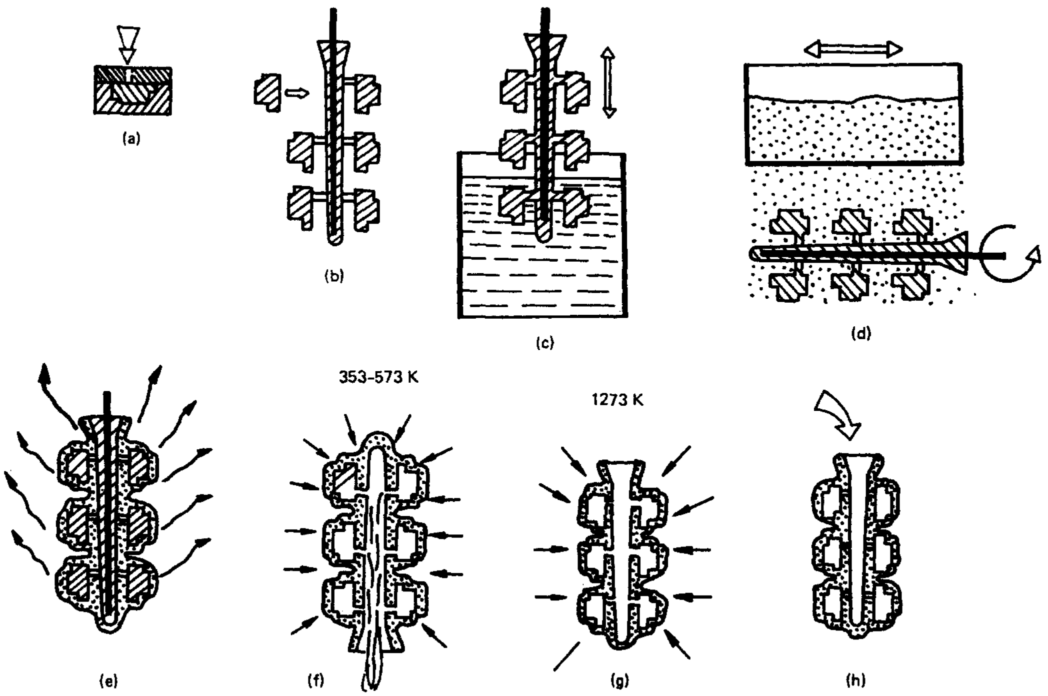


Figure 3.19 Investment casting. (a) Casting of pattern in the chill; (b) collecting of patterns with pouring system; (c) making of the shell in the slurry (up to 8–10 shells); (d) covering each shell with powder; (e) drying or hardening of shells; (f) removal of patterns by melting; (g) annealing; (h) pouring

Table 3.7 Coating mixtures used in the 'lost-wax' process

Materials	*Application*
94 parts of 325-mesh silica, 56 parts of 325-mesh alumina, 37 parts of 40-mesh silica, 80 parts of liquid (4 parts of sodium silicate, 1 part of 2% polyvinyl alcohol)	High-melting-point alloys
90% silica, 6% magnesia, 3% ammonium phosphate, 1% sodium phosphate, water or 1% hydrochloric acid	Cast steels
52% SiO_2, 45% $Al_2O_3 + TiO_2$, 0.87% Fe_2O_3 Molichite*	

* From English China Clays Sales Ltd.

The moulds are expendable, and are made using permanent patterns. This process is particularly suitable for the manufacture of castings made of high-melting-point alloys such as cast steel. The dimensional accuracy of steel castings of dimensions up to 25 mm is ±0.08 mm, whilst for dimensions above 370 mm the accuracy is ±1.14 mm.

Table 3.8 Second and subsequent coatings used in the 'lost wax' process

95% sand, 5% alumina cement; then 30% water additionally
90.6% sand, 7.1% calcium phosphate, 2.3% MgO; then 51% water additionally
1460 g of clay/silica flour (3 : 17), 800 ml of ethyl silicate (No. 40)/denaturated ethyl alcohol/HCl in water (37.5 : 60 : 2.5)

The moulding mixture is composed of a fine-grained material (e.g. MgO, $MgO{\cdot}Al_2O_3$ or $ZrO_2{\cdot}SiO_2$), a binder (e.g. hydrolysed ethyl silicate) and a hardener, which is a gelling agent. The hardener causes a change in the consistency of the moulding sand: it changes from a dense liquid to a plastic composition and then to a ceramic shell of high strength at the final stage.

Moulds are made by the following techniques.

(1) *Block (monolith) moulds*—the whole mould is made of the Shaw composition; these moulds are used in unit production; they are expensive.
(2) *Two-layer moulds*—the mould is made of two types of the Shaw composition; these moulds are used for art castings.

(3) *Shell moulds*—these are manufactured by means of pressing; they are used for medium-sized castings in an automated production cycle.
(4) *Metallic moulds*—these are cast iron moulds lined with the Shaw composition.
(5) *Composite moulds*—one layer is made of the Shaw composition and another layer is made of a filler (i.e. a mixture of chamotte and water glass); this method enables savings of up to 60% total costs.

The ceramic material used to prepare the Shaw composition is usually mullite ($3Al_2O_3{\cdot}2SiO_2$) or zirconium silicate ($ZrO_2{\cdot}SiO_2$). The Shaw composition is characterised by high refractoriness, low expansion, absence of molten alloy, mould reaction, and low price.

The binder is hydrolysed ethyl silicate of pH 1.5–3.0. Alcohol, water and acid are used for hydrolysis. The acid is usually concentrated HCl; sulphuric or orthophosphoric acids are occasionally used.

The gelling material is an ammonium salt or carbonate. The binding rate of the Shaw composition depends on the amount and type of hardener introduced into the mixture of binder and ceramic material.

The process of making two-layer moulds is shown in *Figure 3.20*. Two patterns are used here. One is an oversize pattern made of wood or plastic and characterised by a low grade of accuracy. It is 8–10 mm larger than the mould, i.e. it is larger by the thickness of the facing sand layer made of the Shaw composition. After placing this pattern on a plate, the box is filled with filler (e.g. a chamotte mixture) and hardened with CO_2.

After hardening of the mould, the oversize pattern is removed, and the mould is placed on another pattern plate with a pattern exactly reproducing the shape of a casting. This pattern determines the accuracy of the casting and, therefore, must be made with great precision. The mould cavity is then filled with the Shaw slurry. After hardening, the mould is turned over, the pattern lifted off and the alcohol is burnt out from the Shaw composition. During this burning out process water is removed from $Si(OH)_4$ and the sand loses its elasticity, becomes rigid and acquires proper strength. If burning out begins immediately after lifting off the pattern, the surface of the mould is covered with small grazes. These are small enough to keep the molten metal from penetrating them, but they increase mould permeability and its resistance to temperature variations. Further increase in the mould strength and stability takes place during baking at temperatures of 1173–1273 K.

If the mould is not burnt, the surface of the mould cavity is hardened using a special mixture. If burning is not applied, volatile matter is not removed from the mould and microcracks are not formed. Hence, with no burning, mould permeability is lower and the mould is suitable for the manufacture of cores.

3.5 Shell moulding

This is a variation of investment moulding which can be fully mechanised. The process, invented by J. Croning in 1944, is used very widely because of its high dimensional accuracy ($\pm$0.003 mm). This accuracy is due to the use of pattern plates with stabilised dimensional deviations ($\leq$0.025 mm). Other advantages of this process include:

(1) reduced consumption of moulding materials;
(2) higher productivity per square metre of foundry shop;
(3) long bench life of moulds;
(4) reduced costs of knocking out, fettling and labour; and
(5) elimination of machining.

The disadvantages of the shell moulding process are the high cost of pattern plates and resin, and limited weight of possible castings.

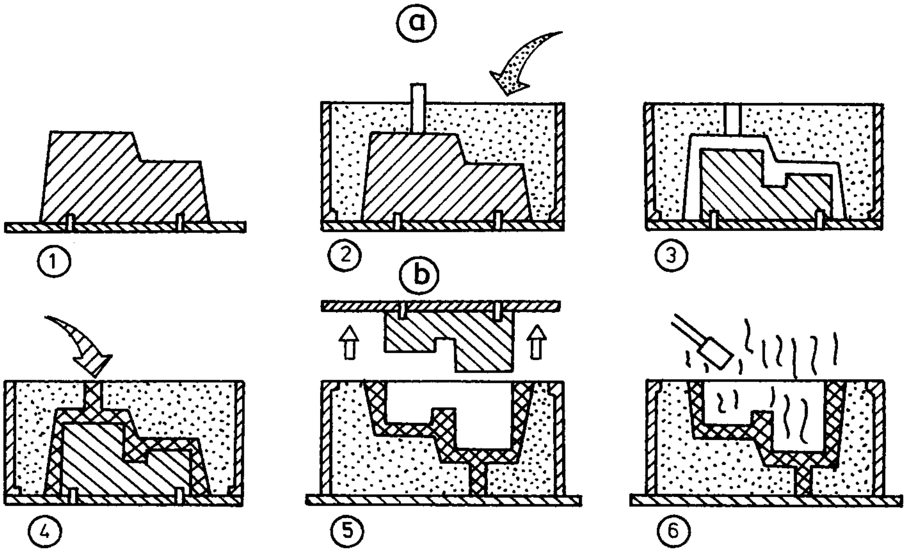


Figure 3.20 The making of two-layer moulds. (1) Enlarged pattern; (2) making of the mould with backing sand; (3) proper pattern; (4) facing sand (with ethyl silicate); (5) drawing of pattern; (6) drying. (a) First stage: making the mould; (b) second stage: making the proper shell

The process is, therefore, applied in large production runs in the automotive and precision parts industries.

In the Croning process, the main materials used are washed silica sand (97% silica), resin (phenolic and aldehyde) and a hardener (usually hexamine).

After mixing the sand and resin the process of hardening occurs only after heating the sand to a certain temperature. Thus it is possible to store the sand in the place where it was prepared and then to carry it to the moulding stand.

The sand used in this process should have a grain diameter of 0.1—0.15 mm and a sintering point of 1620 K. The sand may contain up to 0.5% iron oxides and up to 0.5% carbonates.

The binder for the sand is a resin which, during the formation of a shell, will first form a thin layer and then undergo hardening. Resins used for this purpose are furfuryl or polyester (saturated or unsaturated) resins.

Shell moulds are made in the following way. The pattern plates are coated with a material which facilitates the separation of the shell. These are usually silicon oils, wax or kerosene.

The steps involved in the shell moulding process consist of:

(1) coating the pattern plate or core box with a parting agent;
(2) heating the plate to 500–530 K; and
(3) pouring the sand onto the plate and waiting until a 5–8 mm thick shell is formed.

The excess sand is then removed and the shell heated to 620 K for further hardening. Both mould halves prepared in this way are joined together, either mechanically or with a glue, placed in a container and surrounded with coarse sand. At this stage the mould is ready for pouring. Similarly, cores can also be made as a solid or in halves.

A typical machine for making shell moulds is shown in *Figure 3.21*. The machine comprises a blower fitted with a membrane squeezing system. The resin-coated sand (1) is introduced under pressure through ducts (2) into a space between the pattern plate (3) and the contour plate (5). Attached to the contour plate is an elastic membrane (4). The sand is blown into cavity B, and then the compressed air is blown into cavity A. The air presses the sand to the pattern plate through the membrane. The space between the heated pattern plate and the box is sealed with a rubber which is resistant to the prevailing temperature.

Sand with 3.5% resin is fed from the tank (8). The blowing time is 1–1.5 s at a pressure of 6×10^5 Pa. After filling the box, squeezing is done for 5 s at 1.5×10^5 Pa pressure.

Another technique of making shell moulds is the hot-box process. This process consists of shooting or blowing the sand mixture into a hot core box (480–500 K) and holding it there until hardening takes place.

The core boxes used in the hot-box process are made of cast iron or cast steel; they are heated either electrically or with gas. The heating elements are a part of the shooting machines and their operation is automatically controlled. After heating to 450–550 K the core box is coated with a parting agent (water emulsion of silicon oils, or wax), and the sand is then shot inside. After hardening of the external layer of sand the cores are removed automatically.

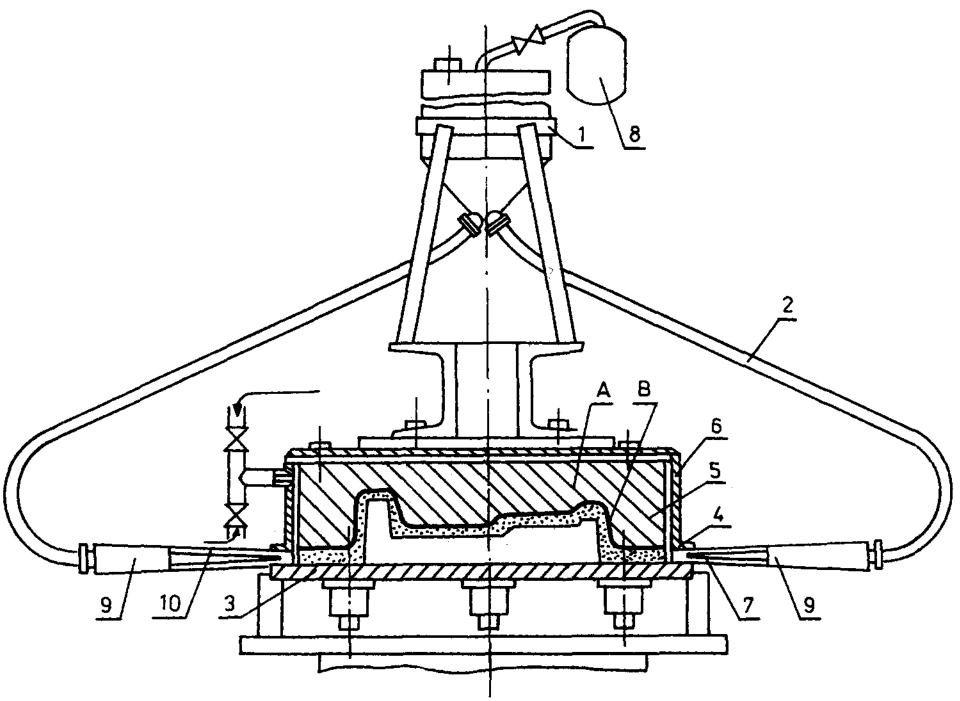


Figure 3.21 Machine for making shell moulds. 1, Resin coated sand; 2, duct; 3, pattern plate; 4, elastic membrane; 5, contour plate; 6, bottom flask; 7, nozzle outlet; 8, reservoir with resin coated sand; 9, nozzle; 10, air valve. A and B, cavities. (Reproduced by permission from Dańko[12])

3.6 Sintering

As one of the processes of powder metallurgy (powder production, mixing and blending, compaction, sintering and testing), sintering has found wide application in the manufacture of new types of material.

The use of powder metallurgy has enabled the fabrication of alloys which cannot be produced by melting and casting and the development of high-strength materials, composites, refractory compounds, and dispersion strengthened metals.

The final properties of a sintered material depend on the properties of the materials used for the manufacture of the powders and on the properties of the mass powder. The latter are determined by structure, the density, the melting point of the particles, the average size and shape of the particles, the particle size distribution, the specific surface of the particles, and the friction conditions.

The powders are manufactured by milling, grinding and crushing of solids. If the metals are in the liquid state, powder is produced by atomisation, i.e. by breaking up the molten stream into gases or liquids. The particles are mixed to obtain the required powder characteristic. The essence of the process is shown in *Figure 3.22*. The desired particle size and size distribution are obtained by crushing, milling (ball or rod) and grinding.

The next stage in powder metallurgy is the mixing of powders. The aim of this process is to introduce the alloying elements, binders (glycerine, camphor, or spirit) or sliding agents (e.g. graphite, oils, wax, or paraffin in amounts of 0.3–1.0%). Alloying elements are added to obtain certain predetermined technological and mechanical properties of the sinters. Sliding agents are used to obtain a more advantageous distribution of pressure during the subsequent process of consolidation. Binders are added when the alloying elements differ in terms of density or particle size from the base powder.

During mixing of the powder changes occur in the particle size (disintegration) and surface structure (oxidation). In addition, during mixing there is also abrasion of the particles due to the presence of the introduced additives and due to the walls of the mixer.

The next stage in powder metallurgy is the consolidation of the powders. Moulding of powders into objects of a given shape is carried out with or without the application of pressure. Without the application of pressure compaction involves unidirectional and isostatic pressing, rolling, forging and explosive compacting. When pressure is applied, compaction also involves slip casting, sintering of powder in the mould and vibrational compacting.

One of the most popular methods of compaction is isostatic pressing of powders. The method is applied in the case of sinters of complicated or oblong shape. In sinters of this type there are some local differences in density.When uniaxial pressing is replaced by isostatic pressing these differences in density are removed.

Isostatic pressing essentially consists of placing an elastic mould (made of neoprene rubber, urethane or polyvinyl chloride) and a powder in a chamber with compressed air or liquid. This causes the same pressure to act simultaneously on all parts of the mould (*Figure 3.23*) and enables the application

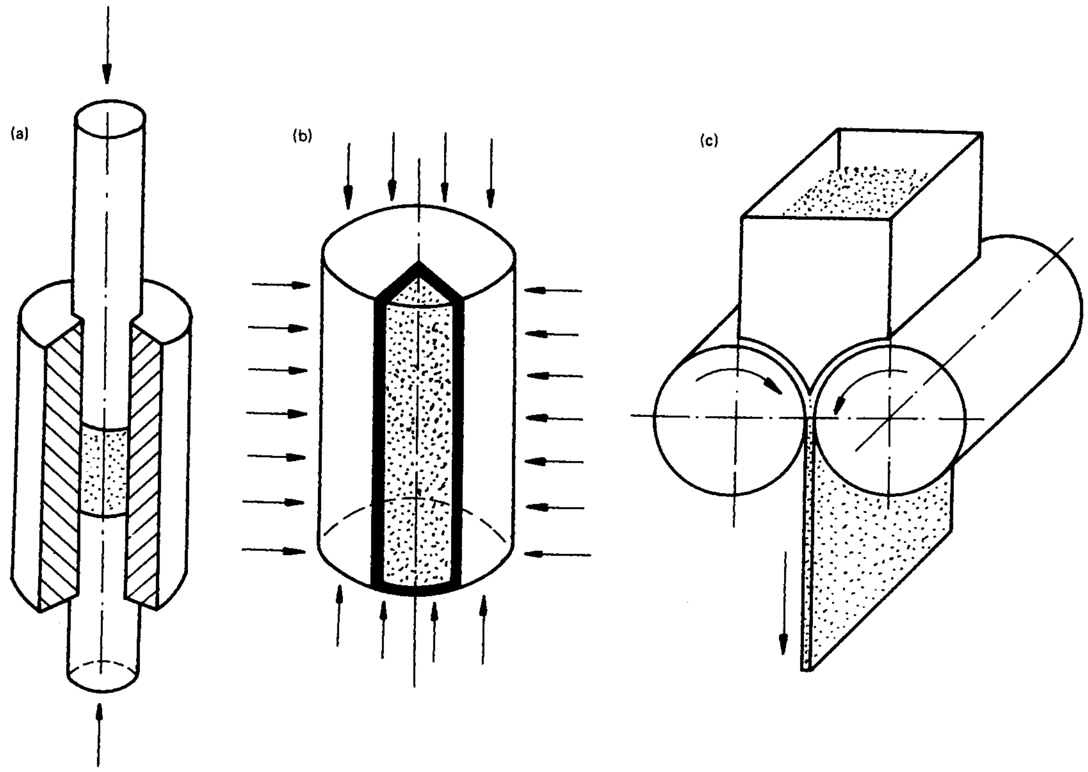


Figure 3.22 Compacting of powder in the sintering process: (a) pressing; (b) in an elastic mould; (c) by rolling

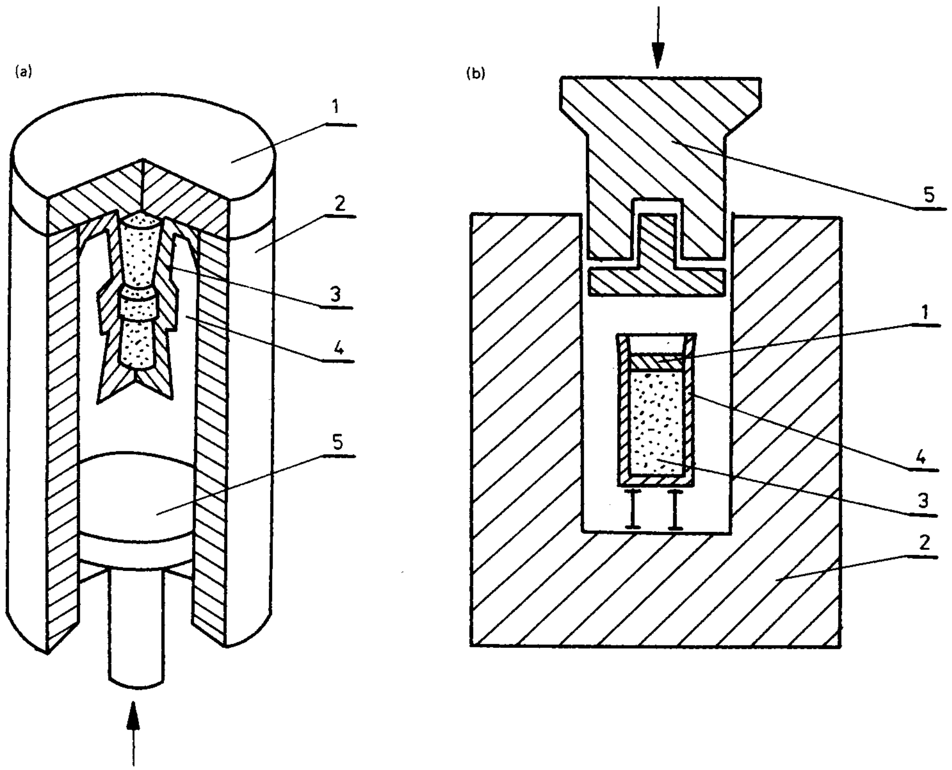


Figure 3.23 Compacting a powder: (a) with stable mould; (b) with stable mould removed. 1, Stopper; 2, press chamber; 3, powder; 4, mould; 5, plunger

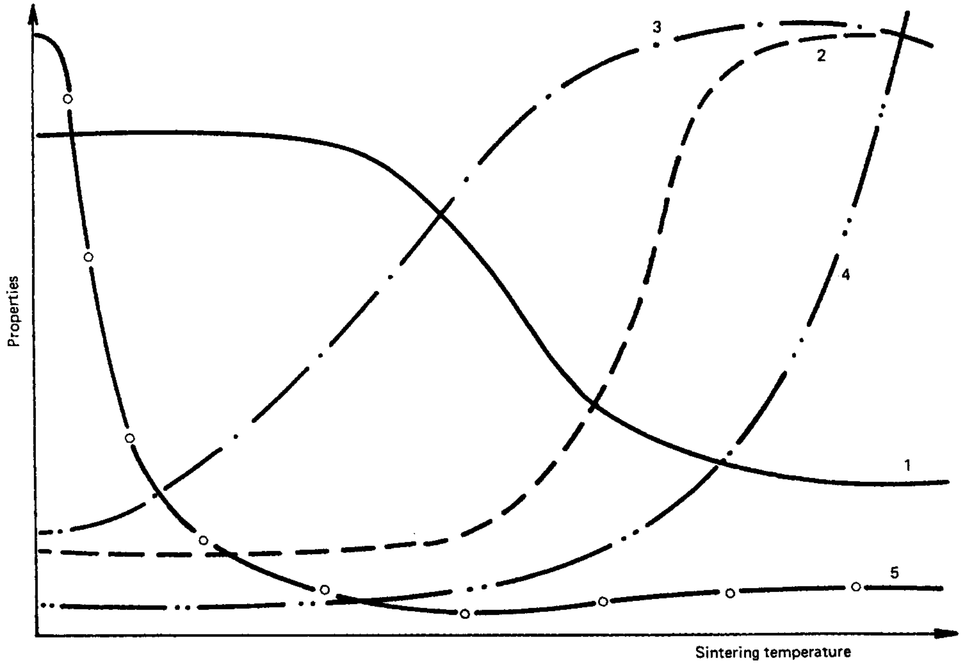


Figure 3.24 The effect of temperature on the properties of powder compacts. (1) Porosity; (2) density; (3) tensile strength; (4) grain size; (5) electrical resistivity

of lower pressing pressures. Owing to these effects, isostatic pressing can be used to consolidate hard and brittle powders as well as to mould large and heavy sintered compacts.

The moulded powder compacts are subjected to temperature in order to strengthen the powder mass. During sintering the powder compacts attain the required properties (*Figure 3.24*).

Sintering is a complex process and is accompanied by phenomena such as:

(1) a change in the surface and crystal lattice at the grain boundaries;
(2) a change in porosity;
(3) recrystallisation;
(4) stress formation; and
(5) changes in foreign constituents (gases and oxide films).

References

1 TRENCKLE, C., *25th International Foundry Congress*, Brussels (1958)
2 CLEGG, A. J., *Foundry Trade Journal*, **3197**, 429, 436 (1980)
3 CHVORINOV, N., *Giesserei*, **74**(14), 461 (1987)
4 WLODAWER, R., *Gelenkte Erstarrung von Stahlguss*, Giesserei-Verlag, Dusseldorf (1961)
5 SKARBINSKI, M., *Poradnik inżyniera odlewnictwo*, WNT, Warsaw (1986)
6 JEANCOLAS, M., *Le Remplissage des Empreintes de Moules en Sable*, Editions Techniques des Industries de la Fonderie, Paris (1966)
7 LEWANDOWSKI, L., 'Sand mixtures' (in Polish), *Publication No. 820*, Academy of Mining and Metallurgy, Cracow (1986)
8 SKARBINSKI, M., *Uruchomienie produkcjiw odlewni*, WNT, Warsaw (1972)
9 GREEN, R. E., *Machinery*, **2528**, 953 (1961)
10 BALEWSKI, A. and DIMOV, T., *The British Foundryman*, **LVIII**(7), 280 (1965)
11 FROMMER, G. and LIEBY, S., *Druckgusstechnik*, Springer-Verlag, Berlin (1965)
12 DANKO, J., 'Particular machines for foundries' (in Polish), *Publication No. 521*, Academy of Mining and Metallurgy, Cracow (1976)
13 LONGA, W., *Risers for Sand Castings* (in Polish), Slask, Katowice (1976)
14 BARANOWSKI, A., HAJKOWSKI, M., IGNASZAK, Z. STODULNY, C. and WASOWICZ, H., *Poradnik inzyniera odlewnictwo*, WNT, Warsaw (1986)
15 GRAF, H., *Stahl u. Eisen*, **96**, 117 (1976)
16 Bühler Catalog, *Technical Information No. 4*, Switzerland (1985)

4 Metal Forming

T Z Blazynski

Contents

4.1 The origin, nature and ultilisation of plastic flow 4/3
4.1.1 Introductory observations 4/3
4.1.2 Classification of processes 4/4
4.1.3 Characteristics of the basic groups of processes 4/6

4.2 Process assessment 4/11
4.2.1 Force/torque estimate 4/11
4.2.2 Flow characteristics 4/12

4.3 An outline of the theory of plasticity 4/13
4.3.1 Basic concepts 4/13
4.3.2 Yielding and criteria of yielding 4/17
4.3.3 Redundancy and geometry factors 4/18
4.3.4 Stress–strain relationships 4/19
4.3.5 Mathematical modelling of metal-forming processes 4/19
4.3.6 Simulation of metal flow 4/23

4.4 Tool design 4/27
4.4.1 Concepts of design 4/27
4.4.2 Lubrication 4/31
4.4.3 Tool materials 4/32
4.4.4 Residual stresses 4/33

4.5 Rolling processes and products 4/35
4.5.1 Classification of processes 4/35
4.5.2 Roll design 4/38
4.5.3 Basic mill layouts 4/40
4.5.4 Rolling of plate, strip and sheet 4/40
4.5.5 Rolling of structural sections 4/42
4.5.6 Rolling of seamless tubing 4/42
4.5.7 Tooling for tube rolling 4/48
4.5.8 Flow forming 4/51
4.5.9 Rolling of machine parts 4/53
4.5.10 Sheet roll bending 4/54
4.5.11 Basic rolling theory 4/54

4.6 Forging operations 4/56
4.6.1 Basic concepts 4/56
4.6.2 An outline of open-, impression- and closed-die operations 4/57
4.6.3 Special applications of closed-die forging 4/62
4.6.4 Subsidiary forging operations 4/65
4.6.5 Rotary forging 4/66
4.6.6 Rotary tube forging 4/66
4.6.7 Orbital forging 4/69
4.6.8 Cast-preform forging 4/69
4.6.9 Isothermal forging 4/70
4.6.10 Tooling in standard forging operations 4/70
4.6.11 Estimation of forging loads 4/72

4.7 Extrusion 4/73
4.7.1 General concepts 4/73
4.7.2 Basic extrusion operations 4/73
4.7.3 Conventional tube extrusion 4/78
4.7.4 Cold extrusion processes 4/80
4.7.5 Continuous extrusion 4/86
4.7.6 Section and sheet extrusion 4/88
4.7.7 Die design and metal flow 4/88
4.7.8 Die materials 4/96

4.8 Cold drawing of wire and tube 4/96
4.8.1 Introduction 4/96
4.8.2 Basic concepts of wire drawing 4/97
4.8.3 Non-standard wire-drawing techniques 4/98
4.8.4 Mechanisms of wire drawing 4/98
4.8.5 The drawing stress 4/99
4.8.6 Tool design in tube drawing 4/101
4.8.7 Special tube-drawing operations 4/104
4.8.8 Drawing loads 4/105

4.9 Sheet-metal forming 4/106
4.9.1 Forming processes 4/106
4.9.2 Shearing operations 4/107
4.9.3 Bending 4/108
4.9.4 Spinning 4/110
4.9.5 Stretch forming 4/110
4.9.6 Deep drawing 4/111
4.9.7 Hydroforming and hydromechanical forming 4/111

4.10 High-energy-rate operations 4/111
4.10.1 Introduction 4/111
4.10.2 Sources of energy 4/113
4.10.3 Forming systems 4/118
4.10.4 Explosive welding 4/121
4.10.5 Powder compaction 4/121
4.10.6 Explosive forging 4/124
4.10.7 Explosive hardening 4/126

4.11 Superplastic and mashy state forming 4/126
4.11.1 Superplasticity 4/126
4.11.2 Mashy state operations 4/127

4.1 The origin, nature and utilisation of plastic flow

4.1.1 Introductory observations

The initiation of plastic flow owns its origin to a variety of micro-effects associated with the application of an external loading system, the level of which cannot be contained within the elastic recovery capabilities of the material. The dimensional change (or changes) becomes permanent and leads to either a desirable accommodation of the metal within the confines of the forming pass or, in extreme conditions, to failure resulting in fracture.

There are two basic mechanisms of flow. Usually, but not always, the initiation of flow takes place when a sufficiently high shearing stress forces the atoms of a crystal to move to a new position of equilibrium. Since an individual atom cannot change its position without affecting the neighbouring structure, the movement will be possible only if a whole layer of atoms moves along. Slipping of the crystalline structure is thus effected and occurs along the planes of highest atomic density. If the load associated with the shearing stress does not exceed a critical value, return to the original position of equilibrium takes place on its removal and the elastic recovery is complete. Generally, stretching is produced by interfacial slip which, in turn, is independent of the value of the normal stress. Slip of the material is made easier by the presence of dislocations or metal-lattice defects. These are instrumental in producing consecutive atom movement which allows plastic deformation to occur.

Smaller plastic deformations can be produced by twinning or by a shift, of constant atomic distance, in a specified direction. A twinned crystal becomes distorted (in a mirror-like fashion) along the twinning plane.

Although, on the microscopic scale, the mechanism of yielding and the subsequent plastic flow are fairly well understood, it is nevertheless very difficult to extrapolate from the micro-condition to the macro-state with which engineers are constantly concerned. It is for this reason that the concept of continuum of matter is generally adopted and the macro-effects are used as criteria of design and general operational purposes.

Even with the acceptance of the continuum approach, it is still clear that a differential in flow is likely to occur. Although most materials are regarded as isotropic, or as continua in which mechanical properties are independent of direction, in a polycrystalline conglomerate lattice faults are usually present and even on this level perfect isotropy is unlikely. Forming of a metal—which takes place prior to the final manufacture of the component—will impose a directionality of flow and a consequent orientation of the grains. Clearly, therefore, directionality in properties must be expected and will be reflected in the level and type of the material response when the specimen is tested in two mutually perpendicular directions. Anisotropy becomes a recognisable property which has to be accommodated in the computation of the relevant numerical values.

A particular case of a micro- rather than macro-anisotropy is that of the *Bauschinger effect* which manifests itself in the reduction in the level of the yield strength of the material on reversal of the direction of plastic deformation imposed.

The Bauschinger effect is associated with the anistropy of the individual grains which, on application of a specific load, find themselves either in the elastic or plastic state. The subsequent removal of the load produces a conglomerate that varies in its properties and remains residually stressed. If the type of stress imposed thereafter corresponds to the prevailing residual, yielding will occur at a lower stress level than the original condition of the material would suggest.

Although the realisation of the presence of these phenomena and the understanding of their effect on plastic deformation are of relatively recent origin, the lack of this knowledge has never, in recorded history, prevented the human race from utilising the empirically established facts.

Metal forming, i.e. changing the shape of the material without actually removing any part of it, was practised at least 3000 years ago in Egypt, where hammer forging to produce gold sheet, cut subsequently to make wire, is recorded in the Bible to have taken place. Rolling in wooden mills was employed to manufacture papyrus. Manual swaging and wire drawing were well established in the Middle Ages but, naturally, were limited in scope by the power available. It was only with the advent of the Industrial Revolution that progress was made and processes like extrusion and cross and longitudinal rolling became available. But even here, the restrictions imposed by the low quality of tool materials, lubrication problems and the lack of understanding of the basic precepts of plasticity impeded progress until, in some cases, well into the twentieth century.

The ever-increasing demand for high quality products—often of sophisticated shape in difficult to process materials—economically produced, fabricated or semi-fabricated, combined with the rising cost of metallic engineering alloys has focused attention on metal-forming processes and techniques.

The emphasis here lies on the 'chipless' approach to shaping. This provides an economical, direct means of converting a cast ingot to slab, plate, billet or bloom and then—in another chipless operation—of changing these basic shapes into profiled finished or semi-finished products. The avoidance of the removal of the material during a forming operation enhances the economics of the process by reducing wastage associated with the swarf-producing machining. Whereas the latter has, of course, a very considerable and necessary role to play in the range of manufacturing activities, its indiscriminate use (a feature of the early years of plentiful supply of cheap labour and materials) is no longer acceptable when high tonnage of accurately manufactured product can be obtained at a much lower cost.

In the most simplistic terms, the desired change in shape is effected either in the cold, warm or hot state (the latter below the melting point of the material) by the application of external forces, pressures or torques of sufficient magnitude to induce plastic flow, and thus a permanent set, of the material through the forming pass. Depending on the operation, the material is forced to flow between driven rolls, through (or into) open or closed dies, or between sets of dies and rolls. Solid or hollow sections are thus produced from the initially solid blocks of metal.

The *standard* basic operations are:

(1) rolling (flat, oblique or longitudinal),
(2) extrusion (axisymmetrical or asymmetrical),
(3) drawing (solid or hollow components),
(4) sheet forming (deep drawing, bending, pressing or bulging),
(5) forging (solid and hollow sections), and
(6) cropping (shearing and piercing).

Within the compass of any of these operations, a number of variants exists which reflects not only a variety of manufacturing routes and subroutes, but also the nature, properties and characteristic responses of the processed materials. Modern metal-forming technology makes use of solid and semi-solid ('mashy' state), and superplastic, as well as explosively pre-

wedled metallic composites and dynamically compacted particulate matter. Mixtures of metallic and/or ceramic and polymeric materials are formed to manufacture composites of very specific properties. The problem of forming these into desirable shapes presents the engineer with new and often difficult situations to solve. Selection of the appropriate forming process, the tool design, the effects of the pass geometry on the final physical and mechanical properties of the product, the dimensional accuracy, and the achievement of the as near as possible final shape in the minimum of operational stages have to be faced.

The apparently simple sequence of ingot–slab–semifabricate–finished product becomes complex unless there is good understanding of the basic characteristics of the individual processes and an appreciation of the principles of the theory of plasticity, as well as that of the concepts of tool and process design. The bases for and fundamentals of the major processes and technological developments are discussed in the following sections of this chapter, but detailed treatment of the individual topics is only indicated by reference to the appropriate literature.

4.1.2 Classification of processes

For a given application, the selection of the correct process necessitates the introduction of a criterion of process classification. Since hot working homogenises and refines the crystallographic structure of the material and thus, ultimately, improves its strength and toughness, whereas cold working increases strength, hardness, dimensional tolerances and improves surface finish, these temperature-induced effects are often used to differentiate between the various manufacturing methods.

Important as the processing temperature is, in some circumstances other criteria of classifying metal-forming processes may well be more appropriate. From a purely manufacturing point of view, quantity and shape may have to be considered, while the likely response of the processed material to the level and/or rate of stressing, as well as the manner of application of the forming load system, may offer a better clue to the desirability or otherwise of using a particular technique or operation.

The parameters that characterise forming operations give rise to the following possible classification systems:

(1) operational temperature (hot, warm or cold forming),
(2) shape effect (bulk or sheet forming),
(3) operational stress system,
(4) operational strain rate,
(5) starting material (ingot, slab, billet, bloom, slurry, or powder).

4.1.2.1 Operational-temperature criterion

The idea behind the subdivision into hot, warm and cold processing of materials is not only to indicate the nature of the operation, but also to draw attention to the plant and ancillary equipment needed, to the level of force parameters required, and to the likely metallurgical response of the processed material.

An outline of this classification scheme, including only the basic operations, is given in *Figure 4.1*.

Starting with a cast ingot, the primary hot operations of flat, billet and slab rolling, and slab forging will produce the starting stage for the secondary, further processing of the slab into plate, billet or a large forging. These, in turn, will form the first step in the manufacturing route of a more sophisticated, profiled product. Hot operations are carried out at elevated temperatures exceeding annealing and normalising ranges and, consequently, yield a hot-finished product showing a relatively low level of flow stress. However, the force parameters required match the mechanical properties of the material and are also relatively low. It follows that the rate of wear of the tooling can be kept at an economical level especially if the lubrication problems are well under control.

To improve the mechanical properties of the product, while at the same time keeping the loading at a moderate level, warm processing is used. Here, the temperatures are well above ambient but, equally, well below the hot-processing range, and usually slightly less than for recrystallisation. The

Table 4.1 Classification of dynamic regimes

Characteristic time (s) *Strain rate* (s^{-1})	10^6 — 10^6 10^{-8} — 10^{-6}	10^2 — 10^0 10^{-4} — 10^{-2}	10^{-2} — 10^{-4} 10^0 — 10^2	10^{-6} 10^4	10^{-8} 10^6
	Creep rates	*Quasi-static rates*	*Intermediate strain rates*	*High strain rates*	*Very high strain rates*
Primary load environment	High or moderate temperatures	Slow deformation rates	Rapid loading or low velocity impact	High velocity impact or loading	Very high velocity or hypervelocity impact
Usual method of loading	Constant load or constant stress machines	Conventional hydraulic or mechanical machines	Fast-acting hydraulic or pneumatic machines, cam plastometers, low impact devices	High velocity impact devices, expanding-ring technique, high-speed metal cutting	Light gas gun or explosively accelerated plate or projectile impact
Dynamic considerations in testing	Strain versus time	Constant-strain-rate test	Machine stiffness wave effects in specimen and testing machine	Elastic–plastic wave propagation	Shock wave propagation, fluid-like behaviour
	← Isothermal →		← Adiabatic →		
	← Inertia forces neglected →		← Inertia forces important →		

increased material ductility is sufficient to reduce the power requirement of the plant. Cold-working conditions are confined to ambient temperature and are characterised by a high energy requirement—necessitated by large operational forces and/or torques—but result in very high quality final product displaying both good dimensional tolerances and mechanical properties.

A rough guide to the temperature ranges can be obtained by considering the operational temperature/melting point ratio. On this scale, hot working takes place when the ratio is >0.6, warm working when the ratio is 0.3–0.5 (the latter corresponds to recrystallisation conditions), and cold when the ratio is <0.3.

4.1.2.2 Shape-effect criterion

The effect of shape reflects the geometry of both the initial and final component and, consequently, the nature of the change imposed on it by the forming operation.

A process in which a component of a relatively small initial surface area/thickness ratio is deformed in such a way that the ratio is increased, is often classed as a 'bulk deformation operation'. On the other hand, the component of an initially high surface area/thickness ratio, shaped in a process which does not impose any change in the thickness but effects shape changes only, is said to be 'sheet formed'. Any change in the thickness of such a component can easily lead to tensile plastic instability and incipient, localised yielding.

Bulk processes are those of rolling, extrusion, forging and solid- and/or hollow-section drawing. Bending, pressing, deep drawing, spinning and shearing are the main sheet-forming operations.

4.1.2.3 Operational-stress system

Because of the inherent severity of many forming processes, particularly the rotary ones, a consideration of the type and property of the induced stress field is of primary importance. The success of the operation may well depend on its compatibility with the properties of the processed material.

The presence of tensile and compressive stress fields results in the appearance of shearing stresses which, in turn, lead to the sliding of molecular planes and, eventually, to the yielding and plastic flow of the metal. Stress systems containing these components are most likely to give rise to plastic flow which, if it is controlled, will produce the desired amount of deformation.

Purely compressive or tensile systems create conditions of hydrostatic pressure in a triaxial field (absence of shear), or produce shearing stresses in uni- and bi-axial conditions. Clearly, since it is the configuration of the individual stress system that is indicative of the type of deformation which can be expected, its assessment prior to choosing a forming system is imperative. These various possibilities are illustrated, diagrammatically in *Figure 4.2*.

As an indication of the incidence of any of the stress systems, the following, non-exhaustive, list can be considered:

Tensile–compressive systems

Biaxial tension/uniaxial compression:

(1) under a roll of a two-roll piercer,
(2) under a roll of a two- and three-roll piercer, and
(3) under a roll in the helical rolling process.

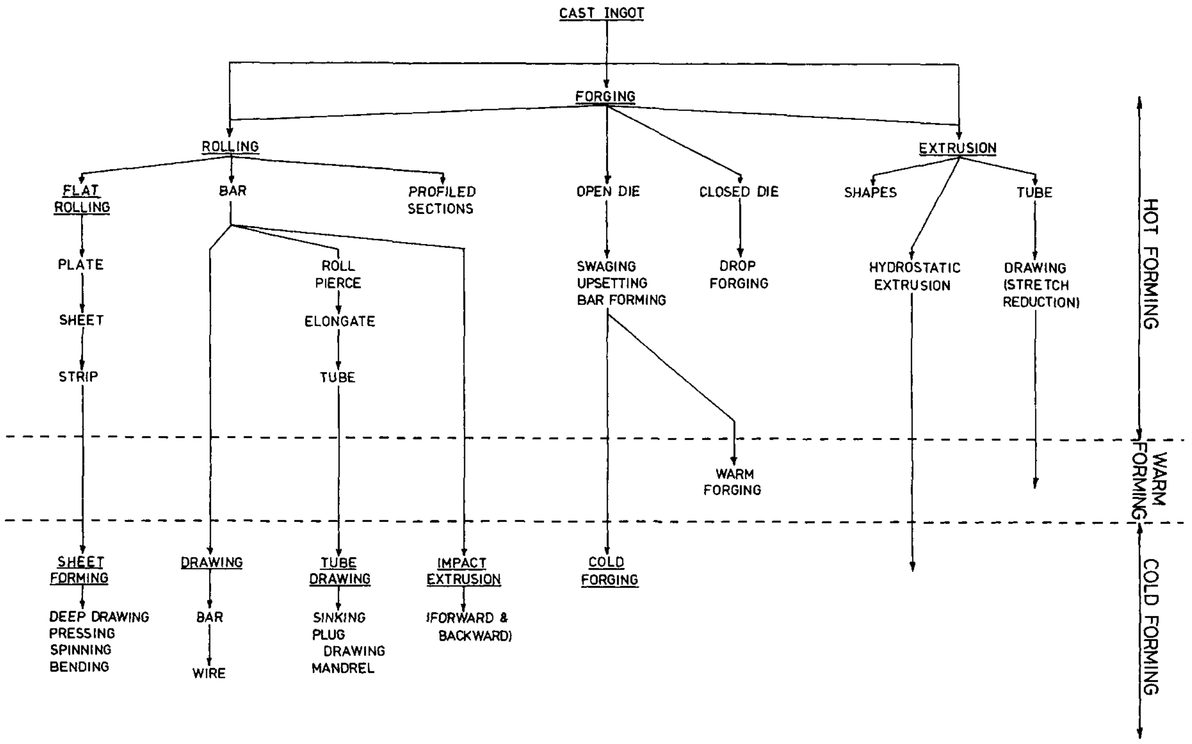


Figure 4.1 Process classification system based on operational temperature

Uniaxial tension/uniaxial compression:
(1) between the rolls in roll forming, and
(2) in the flange in deep drawing.

Uniaxial tension/biaxial compression:
(1) in the drawing die.

Compressive stress systems
Triaxial stresses:
(1) In the oblique zone of a three-roll rotary plug piercing mill,
(2) in the closed forging die,
(3) near the die throat in extrusion of bar, and
(4) under the punch in tube extrusion.

Biaxial stress:
(1) between the rolls of a longitudinal rolling mill with no front and/or back tension, and
(2) in the upsetting, open dies.

Tensile stress system
Biaxial stress:
(1) stretch forming, and
(2) bulging.

4.1.2.4 Operational strain rate

A number of engineering alloys and even some practically 'pure' materials, e.g. commercially pure aluminium, are susceptible to the changes in the rate of straining. Modern technological techniques have either 'speeded up' conventional processes—for instance, wire can be drawn at some 120 m min^{-1} —or have introduced new ones that operate in truly dynamic conditions. Impact extrusion, explosive forming, welding and compaction, and mechanically and electrically induced discharges of energy producing high strain rates, have all combined to introduce an entirely new field of high-energy rate fabrication, known commonly as HERF.

The range of possibilities arising in this context are listed in *Tables 4.1 and 4.2* which provide a detailed insight into the effect of different strain rates and the means of producing them in an industrial environment.

In terms of the mechanical properties of the material, an increase in the operational rate generally leads to a rise in the yield and flow stress and to an accompanying reduction in the value of the strain to fracture. The engineering alloys usually employed fall, in this respect, into two main groups (*Figure 4.3*). One (*Figure 4.3(a)*) will initially display little change in the value of the strain to fracture, but this will be reduced rapidly as the flow stress increases. The other may even show some initial improvement in the strain value, but the deterioration sets in as soon as the optimum strain rate is exceeded (*Figure 4.3(b)*).

Table 4.2 Mass–velocity relationship contributing to total kinetic energy

Forming system	*Mass contribution* (kg)	*Velocity contribution* (m s^{-1})	*Velocity/mass*
Drop hammer	144	122	0.84
Trapper rubber	144	122	0.84
Floating piston	4.54	3050	672
Explosive in gas	0.004	2.75×10^6	687×10^6
Explosive in liquid	0.008	1.52×10^6	190×10^6
Electric discharge	0.008	1.52×10^6	190×10^6

The toughness of these materials will be correspondingly affected (*Figure 4.4*). Consequently, if dynamic working conditions are likely to prevail in the selected process, a classification based on the strain rate may offer a more realistic means of assessment, particularly in the case of cold-forming operations.

4.1.2.5 Starting material

Since some modern processes do not require bulk solids as starting materials, but utilise particular matter and semi-solid substances, a classification based on the initial physical state of the material offers an interesting alternative to the more conventional approach.

Typical examples of unconventional starting materials are: 'mashy' state processing, leading to conventional rolling of composite sandwich components; the Conform-type extrusion, starting with a powder or granulated material, or an explosive compaction of powders.

4.1.3 Characteristics of the basic groups of processes

Of the major processes listed in Section 4.1.1, forging is the most diverse and cannot therefore be described in more general terms. For this reason, the basic characteristics of only four groups of processes are indicated here and those of forging, sheet forming, cropping, etc., are discussed later.

All rolling processes rely on the forces transmitted through the rolls to the material to effect deformation and on the rigidity of the roll system for the dimensional accuracy of the product.

Sheet and plate are initially obtained from a slab by rolling the slab in a relatively simple system (*Figure 4.5*). Driven rolls introduce the material into the roll gap, or working zone of the pass, and reduce the thickness. The success of any further processing to obtain strip rather than a sheet or large area of plate, depends on the ability of the system to maintain a constant width of the processed metal and on reduction of the thickness (this being equivalent to the reduction in the cross-sectional area). These requirements call for a plane strain operation which is possible only if the lubrication of the pass is very efficient. Processing in this mode can proceed in either cold or hot conditions.

A much more complex rolling system is that of longitudinal rolling, which is employed in the production of axisymmetrical billets, bars and hollows (*Figure 4.6*). A train of suitably shaped rolls, mounted on stands (either in pairs or in three-roll configurations) inclined at right angles (between the successive stands) is used, as shown diagrammatically in *Figure 4.6(a)*. A gradual reduction in the cross-sectional area of the material takes place (*Figure 4.6(b)*) as the specimen moves axially forward through the sets of driven rolls. While fully engaged in the train, the processed material experiences, additionally, axial tensions resulting from a differential distribution of successive stand velocities. The ovality of the early passes is slowly reduced along the train until the last stand is reached. Here, the final, circular cross-section is expected to be achieved.

An alternative to longitudinal rolling is offered by the oblique-rolling system in which a single set of two or three driven rolls produces tractive, frictional forces which propel the specimen axially while, at the same time, causing it to rotate. The motion of an element of the worked material is thus forward, but helical.

Figure 4.7 illustrates, using an example of tube rolling, the basic principle involved. In this case, three profiled, driven rolls, disposed at 120° to each other, and inclined at an angle α

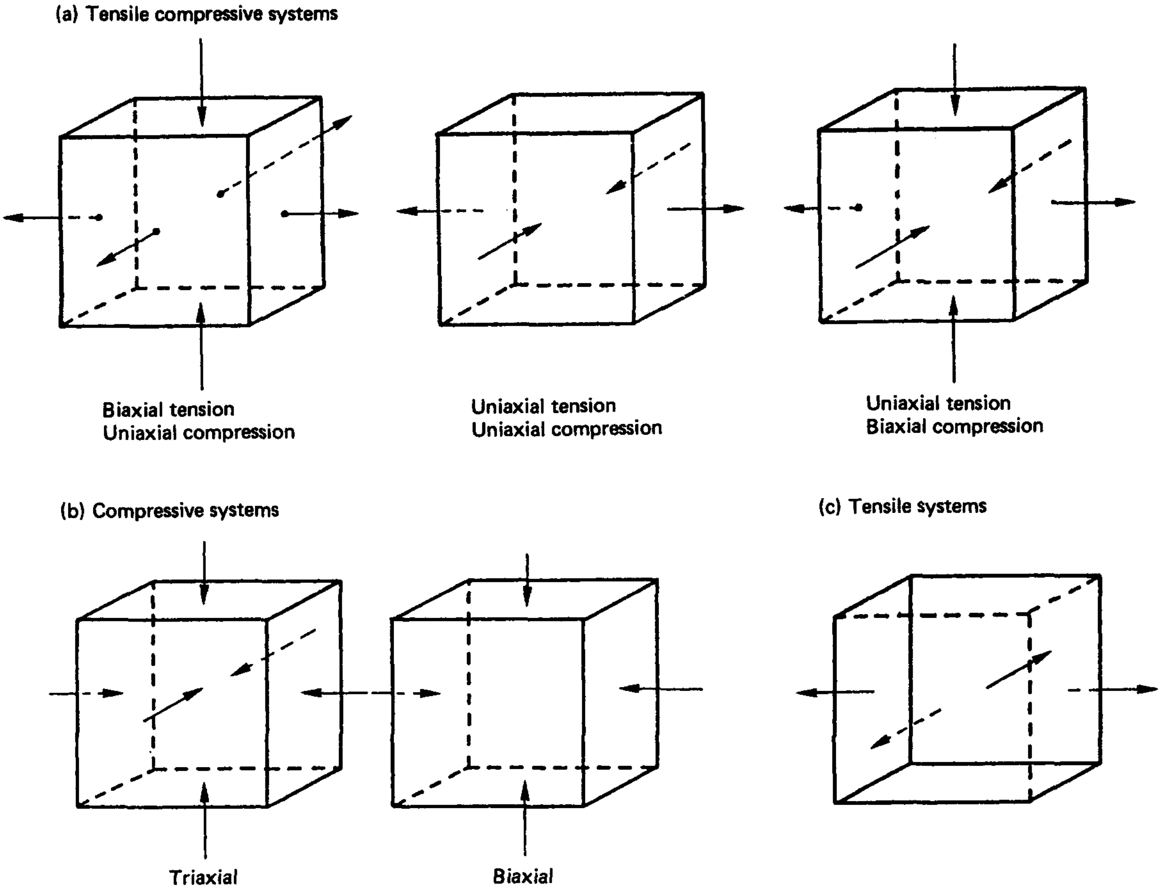


Figure 4.2 Process stress classification system

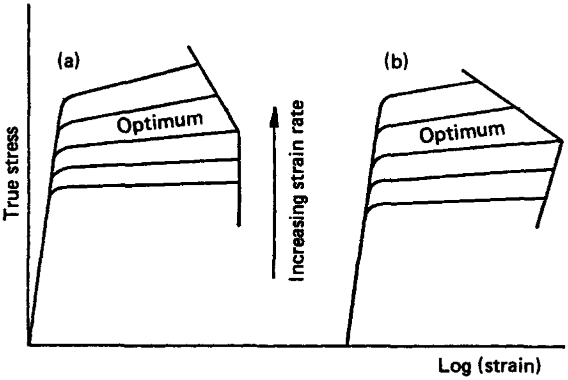


Figure 4.3 Stress–strain (to fracture) relationships

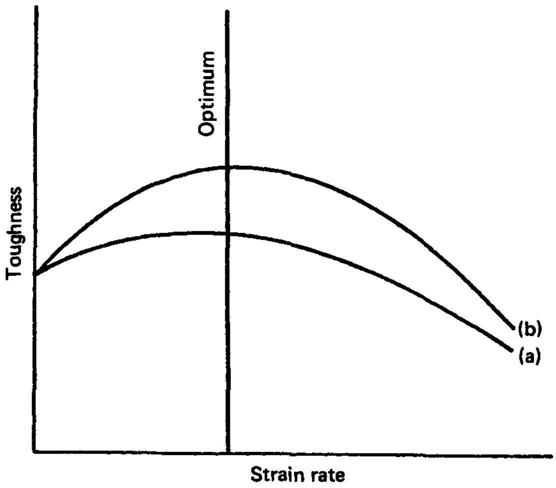


Figure 4.4 Variation in toughness with strain rate for the alloys in *Figure 4.3*

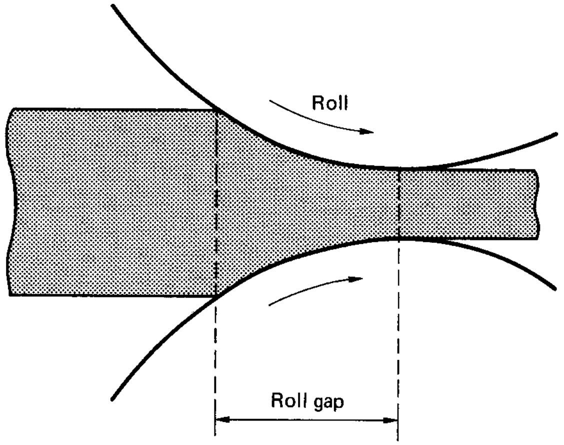


Figure 4.5 The principle of sheet and plate rolling

(the feed angle) to the horizontal mill axis, and an angle β (the cone angle) in the vertical plane, introduce the bloom (supported internally in the bore by a mandrel) into the forming pass. The bloom is 'sunk' onto the mandrel in the zone AB and has its wall thickness reduced on the roll 'hump' BC. Slight elastic recovery takes place along DE. The bloom is thus elongated and its wall is thinned. The amount of deformation imposed depends on the size of the inter-roll opening or the 'gorge'.

These basic characteristics of oblique rolling operations (the variants of which are discussed later) are common to all operations, as indicated, for instance, in *Figure 4.8*. This shows, diagrammatically, the operation of the so-called 'secondary piercing', or 'oblique plug rolling' of a tube — a process in which a long cylindrical mandrel is replaced by a short profiled plug.

On the other hand, processes of profiling by rolling can take various forms, two of which are indicated in *Figures 4.9* and *4.10*. A stepped shaft, required to acquire a series of specific profiles, can be manufactured by oblique rolling in a single three-roll stand (*Figure 4.9*). An operation in which the billet is rotated and fed through a system of driven rolls produces this effect. In another variant of oblique rolling, a two-roll system of helically ribbed rolls (*Figure 4.10*) will produce metal balls out of a solid cylindrical billet.

These few examples illustrate the versatility of rolling operations, a more full discussion of which is given in Section 4.5.

When the initial shape of the work piece has been imposed on it by one of the processes described above, there often arises the problem of how to achieve a degree of further deformation leading, possibly, to the final product. Drawing processes answer this need by providing a means of producing either solid (bar, rod or wire) or hollow tubular sections, either circular or non-circular in shape. The drawing operation is carried out in a die—or a set of consecutive dies forming a tandem drawing system—into which the work piece, with a swaged leading end, is introduced (*Figure 4.11*). An axial force is applied through a gripping device (as indicated by the arrow in *Figure 4.11*) and the work piece is pulled through the die. In the case of a solid specimen, the outer dimension only

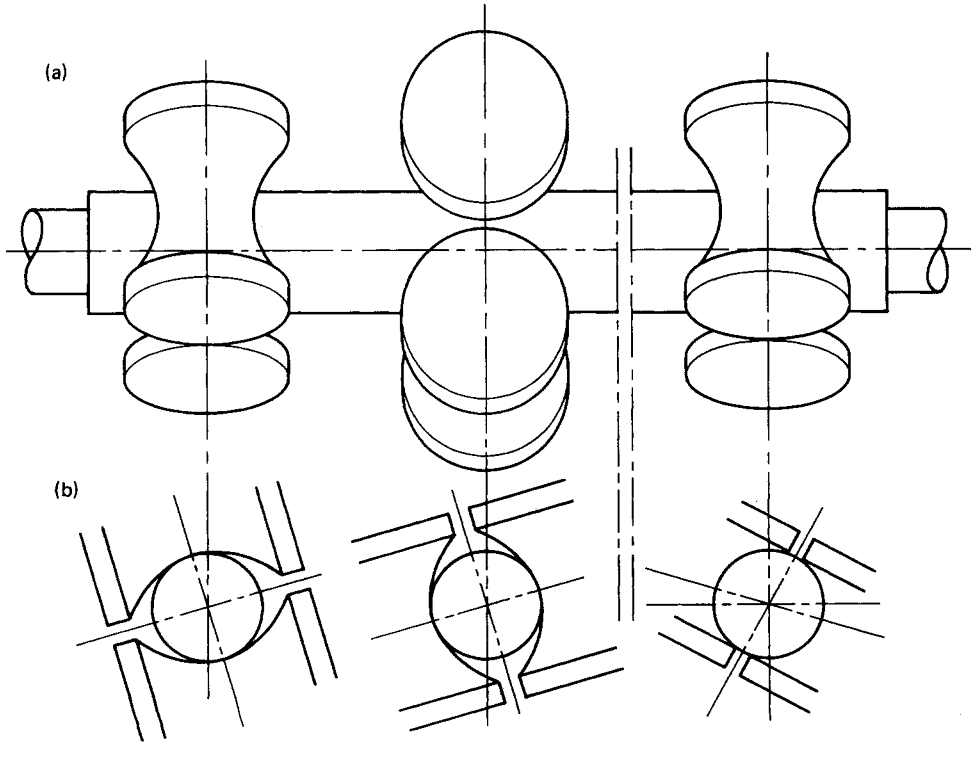


Figure 4.6 Longitudinal rolling: (a) a roll train; (b) successive roll passes

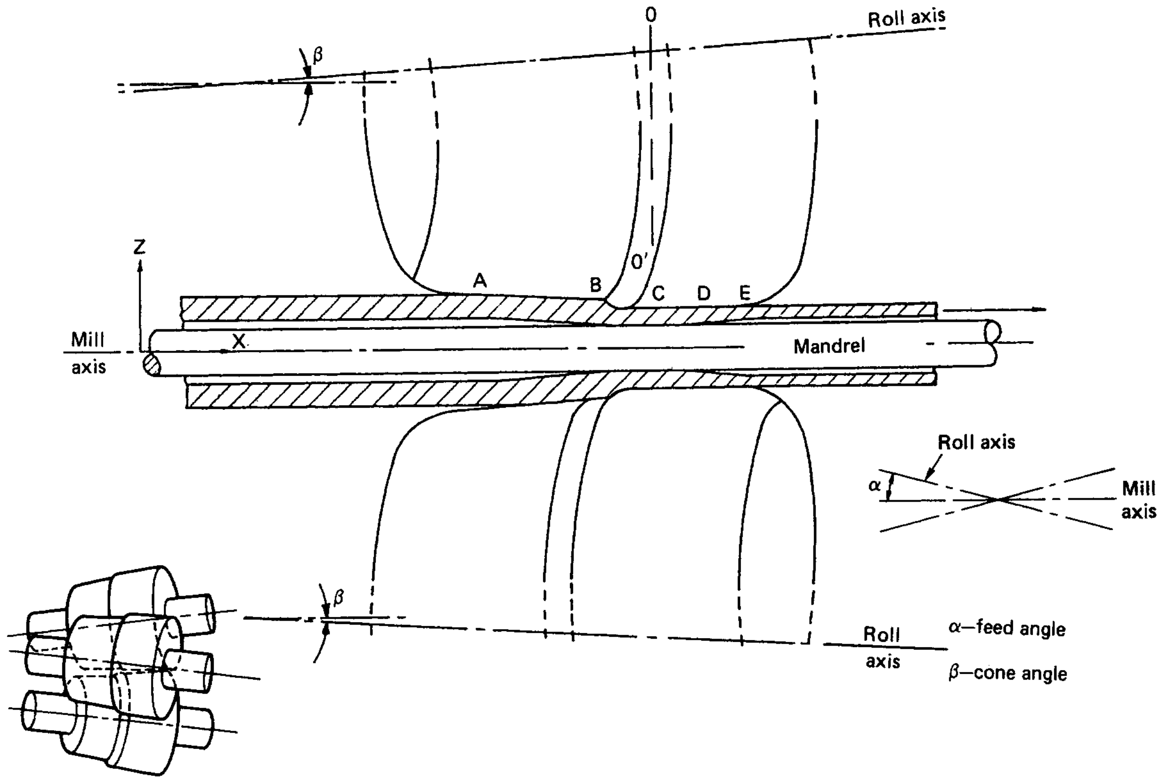


Figure 4.7 Oblique, three-roll tube rolling on a mandrel

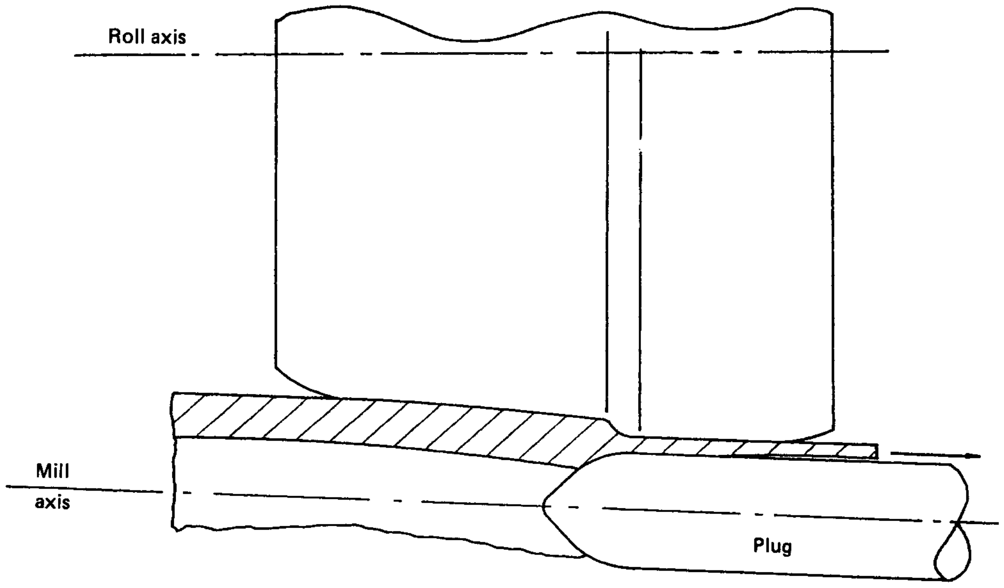


Figure 4.8 Oblique, tube rolling on a plug (secondary piercing)

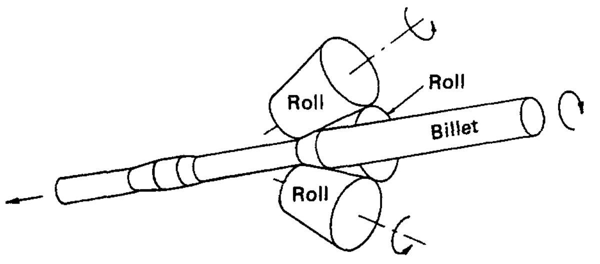


Figure 4.9 Three-roll shaft shaping

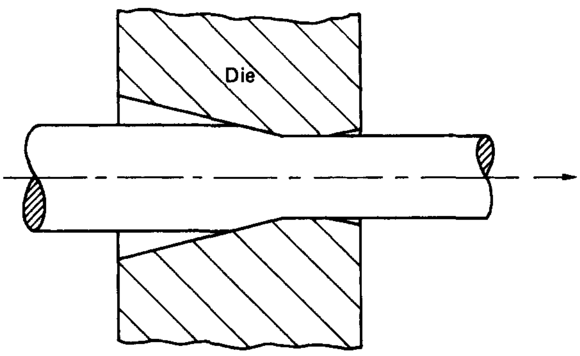


Figure 4.11 Open die drawing

is reduced, whereas with a hollow section there is also a change in the wall thickness. Lubrication of the working zone of the pass (the part of the die surface along which the deformation is effected) is of importance from the point of view of the magnitude of both the drawing load and the induced drawing stress, and in view of the surface finish.

Similar results can be obtained in extrusion, a process in which the starting billet (sometimes referred to as the 'slug') is inserted into a cylindrical container and is then pushed mechanically through a suitably profiled die (*Figure 4.12*). there is a number of variants of this process (see Section 4.7), but the two basic operations are those of forward (or direct) and inverse (or backward) extrusion. In the forward extrusion a solid moving ram is brought into direct contact with the billet and activates the latter by moving it axially forward through the die. In inverse extrusion a hollow ram is in contact with a movable die which bears onto the billet, firmly held in the container. When the pressure exerted by the tooling is sufficiently high to exceed the yield stress of the material, plastic flow is initiated and backward extrusion into and through the hollow ram takes place.

Considerable control over the dimensional accuracy can be exercised in such systems but, again, solution of the lubrication problem is of importance. In this latter context, hydrostatic extrusion (to be described later) provides an important alternative to the conventional arrangements indicated here.

A large group of 'unorthodox', dynamic processes introduces a number of new elements and opens new operational possibilities of using materials which are sometimes difficult to process and of reducing manufacturing costs by dispensing with heavy plant and equipment.

The high-energy-rate processes stem essentially from the usually overlooked fact that the working of metal requires energy and not merely the application of force, and that, in addition, the rate of dissipation of energy is of importance. A simple consideration of the basic equation for kinetic energy shows that a comparatively small change in the velocity of a body will have a more pronounced effect than will a change in

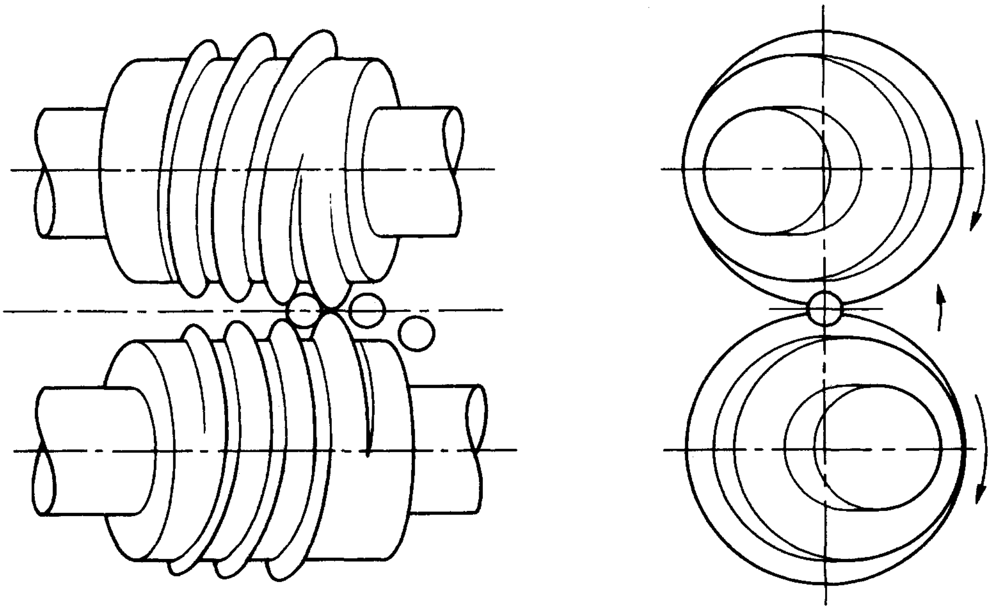


Figure 4.10 Two-roll ball manufacture

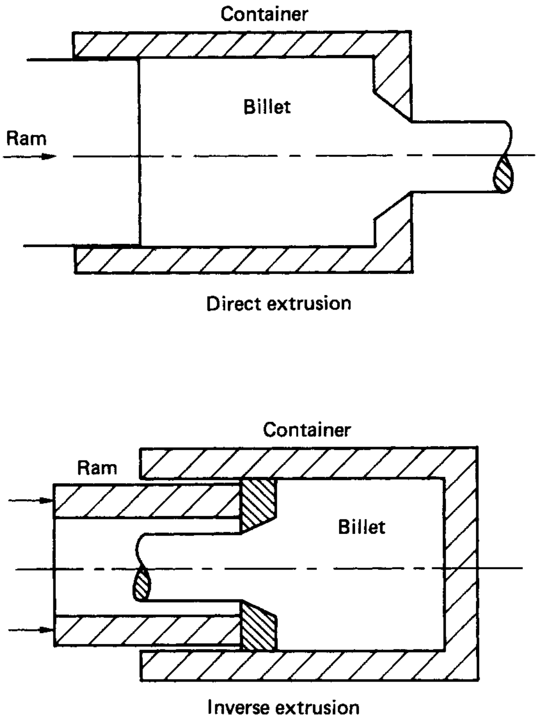


Figure 4.12 Direct and inverse extrusion

its mass. A typical conventional system approaching the conditions of high-energy forming, i.e. drop-hammer forming, is limited in its usefulness by the necessity of using large masses and, therefore, unwieldy and costly equipment.

In a high-velocity system, the mass contribution to the available energy can be negligible, as indicated in *Table 4.2*. The forces generated in a system of this type are easily accommodated by the inertia of the forming dies.

The sources of energy used in the high-velocity systems are chemical explosives, electrostatic and magnetic fields, and pneumatic–mechanical devices.

The basic processes are those of forming (shaping), welding and powder/particulate-matter compaction. A variety of forming systems exists, each displaying specific characteristics associated with either sheet or tube forming, for which it is intended. Consequently, a more detailed description of such processes is given later, but to appreciate the principles involved, an example (*Figure 4.13*) which refers to the explosive bulging of a tube is used here.

A stainless-steel (in this case) tube to be shaped is inserted into a profiled, vented die which is then closed. An explosive, linear charge, capable of delivering the required level of energy, is then placed axially in the tube and either air or water is used as the energy-transmitting medium. On detonation of the charge, high-pressure, high-temperature gases are generated and these propel the tube radially outwards, until it fills the die and acquires the required shape. The two-part die is then opened and the product extracted. The finish of the die surface is of importance because the outer surface of the tube comes into intimate contact with it and will bear the imprint of any machining, etc., marks.

The action of electrical discharge, whether capacitance in type or magnetic, is similar, but the forming system employed will, of necessity, be more complex and sophisticated. Because of the inherent cost of electrical equipment of this type, only relatively small sized components can be formed economically.

4.2 Process assessment

4.2.1 Force/torque estimate

The efficiency of a metal-forming operation can be assessed on two different levels. From the point of view of plant design and its subsequent rational utilisation, it is necessary to have a means of predicting the values of the externally applied forces and torques. This calls for the development of a suitable mathematical model of the operation and for a purely theoretical approach to the determination of numerical values. A theory of plasticity and methods of modelling based on it must be established. These are reviewed in Section 4.3.

However, the characteristics of metal flow, i.e. its incidence and pattern, cannot be neglected in an assessment because they reflect both the chosen geometry of the pass and the

Table 4.3 Summary of the characteristics of laboratory stress–strain testing methods

	Tension	*Simple compression*	*Plane compression*	*Torsion (solid test piece)*	*Torsion (hollow test piece)*
Definability of stress conditions	Not definable. Dependent on metal at high strain	Fairly definable, but dependent on friction at high strain	Definable (plane strain)	Not definable when axial stresses present	Definable (simple shear)
Amenability of features of test to correction	Corrections doubtful, complex measurements (contours) needed	Sound correction methods, but multiple tests needed	No correction needed	Strain rate corrections needed for very strain-rate sensitive materials	—
Information at high strains	Stress–strain curve doubtful	Information below 2.3 strain only	Information below 2.3 plane strain only	No limit*	Information obtained at strains >6
Ductility indication	Obtained	None	None	Obtained	Obtained
Suitability for tests at high temperatures	Suitable	Suitable	Doubtful	Suitable	Suitable
Suitability for tests at high strain rates	Suitable, but some severe experimental difficulties	Suitable	Unsuitable owing to need to change tools	Suitable	Suitable

* Results obtained by the present author up to strains of 10 000.

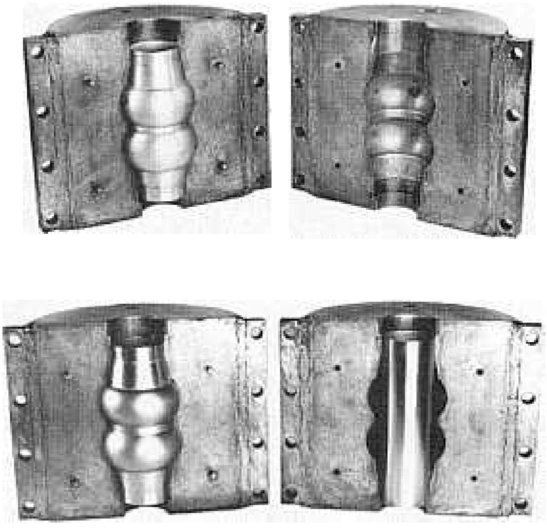


Figure 4.13 Explosive forming of a profiled 18/8 stainless-steel tube in a die

prevailing lubrication conditions, and the response of the processed material to the imposed mechanics of the operation.

The response of a material depends, primarily, on its physical and mechanical properties, but it also depends on the presence, or otherwise, and the type of constraints imposed on the flow. These parameters govern the degree of severity of the operation and, consequently, its efficiency.

Constraints imposed by the geometry of the pass and the initial state of the processed material not only affect the quality of the final product, but also influence the magnitude of the operational force parameters. The economics of the operation will thus be affected from the point of view of both the power consumption and tool wear, and the volume of imperfect or even defective product.

Once the expected characteristics of the flow have been anticipated, the forces, stresses and power parameters can be estimated theoretically.

4.2.2 Flow characteristics

The desired change in shape of the work piece, i.e. the basic objective of the metal-forming operation, normally involves elements of both useful and redundant or unnecessary straining of the material. In this context, useful or homogeneous deformation is associated with the absolutely necessary change in shape, from the initial to the final shape, but not with any other change in the material that does not contribute to this situation. In general, homogeneous deformation is defined by three mutually perpendicular strains and its magnitude and pattern are related to the external forces applied and to the geometry of the system. It does not, however, depend on the mechanical properties of the processed material.

To varying degrees, every forming operation will produce shearing strains associated with the angular and/or longitudinal distortion of the material. Since this type of deformation does not contribute anything to the required change in shape, but merely produces undesirable additional straining, it is clearly redundant to the basic objective of forming.

The pattern of this deformation depends only on the geometry of the system, but its magnitude is governed by the mechanical properties of the material, as reflected by the stress–strain curve, and also frictional effects produced in the working zone of the pass.

In some forming operations, e.g. rotary processes, friction constitutes the only tractive force and its elimination or reduction below a certain critical level results in break down of the operation. Frictional effects can, therefore, be viewed from two different angles. As shears, distorting the structure of the material, they are detrimental to the operation by increasing the magnitude of unnecessary straining, whilst not contributing to the change in shape. As effective tractive forces, they must be tolerated in certain cases in spite of the introduction of additional strain.

Thus the redundant deformation consists of two components: the frictional component, and the deformation produced by the geometry of the pass. The latter manifests itself as shearing at the entry to and exit from drawing dies, longitudinal warping and circumferential twisting in rotary processes, etc. It is mostly the magnitude of the latter component of inhomogeneous deformation that is governed by the mechanical properties of the material and can be assessed from the stress–strain curve. For a given tool–material–lubricant system, the specific contribution of friction depends mostly on temperature and is more pronounced for non-metallic than for the metallic materials.

Detailed study of inhomogeneity calls for the separation of frictional effects from those imposed purely by the geometry of the pass. However, for comparison between, say, the performance of different types of tool profile, this separation is not strictly necessary. If the work piece and tool materials, lubrication and temperature conditions are kept constant, the contribution of friction, although dependent on tool profile, is minimised. The level of inhomogeneous deformation, as measured or calculated on this basis, reflects the performance of the tools and serves as a measure of their efficiency in changing the shape of the work piece.

Typical examples of the incidence and likely pattern of inhomogeneous deformation are given in *Figures 4.14* and *4.15*.

The former refers to the rotary piercing of a solid billet in an oblique rolling operation (similar to those in *Figures 4.7* and *4.8*, but not starting with a hollow). If, using a model material that is capable of being marked internally, the operation is stopped before the whole of the billet has been pierced, and the product is sectioned, the angular distortion along the pass becomes obvious. In the absence of the redundant shear, the originally vertical and horizontal banks or markers retain their angular positions and only change dimensions (homogeneous strains). The distortion seen here is due to a differential shearing of notionally, infinitesimally thin cylindrical elements with respect to each other in tangential directions.

In non-rotary drawing or extrusion processes (*Figure 4.15*), an internally marked billet (*Figure 4.15(a)*) shows that distortion takes place in the axial direction; the distortion can, however, be quite severe. If the distortion is neglected when estimating the strain for the purpose of calculating extrusion pressures or drawing loads, considerable error can be introduced. For instance, if the evaluation of redundancy is based on the distortion of the marker AB (*Figure 4.15(b)*) to its steady-state position A'B', then it could be concluded that the longitudinal shear strain is practically uniform across the section and may be represented by, say, $A_1A_2 = B_1B_2$. If, however, the variation in strain along a number of flow lines, such as aa', bb', etc., is considered, the distribution is non-uniform and corresponds to, say, A_3B_3.

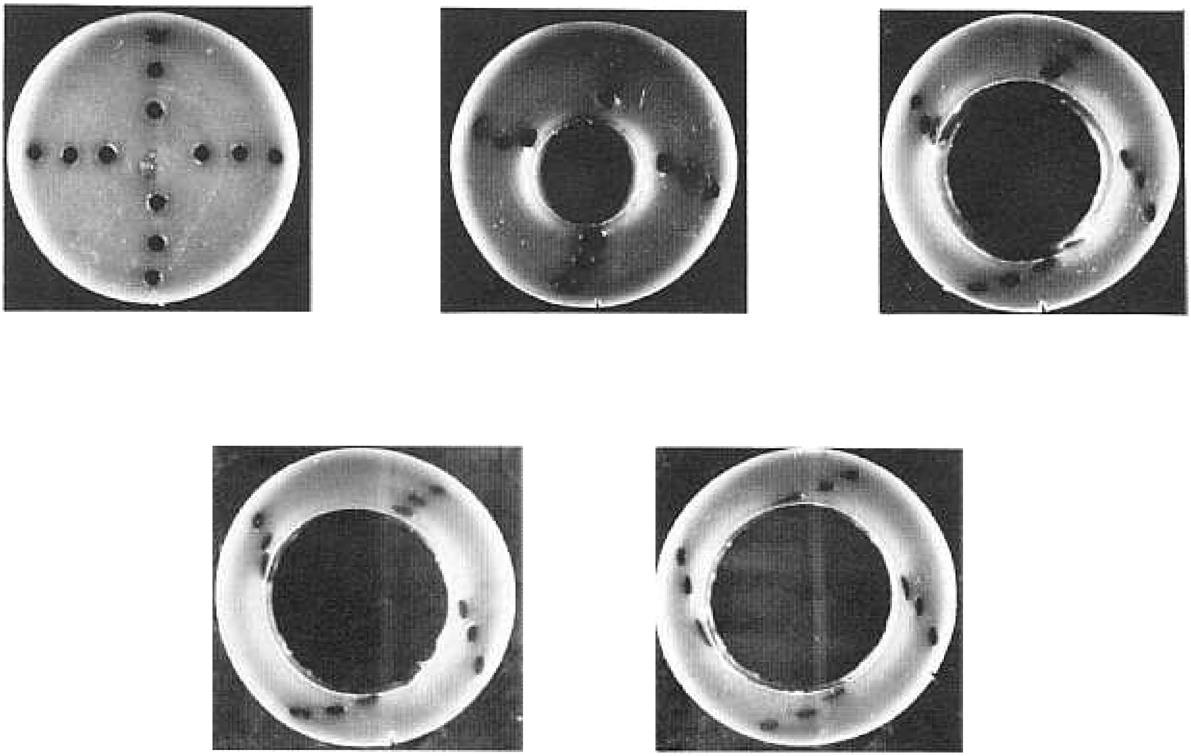


Figure 4.14 Development of redundant, shearing circumferential distortion and distortion due to twist strain in the rotary piercing of a billet

Clearly, the material experiences different levels of straining in different internal sections, and its potential suceptibility to overstraining and, possibly, fracture must be taken into account when designing the pass.

The effect of the additional, redundant shearing strain on the response of the product depends on the type of operation. In cold processes, the presence of shear increases the level of flow stress and, therefore, tends to introduce an element of brittleness into the material. This effect is indicated in *Figure 4.16*. The curve OA′B′C represents the stress–strain relationship in the original, unworked metal. When this material is cold formed and the homogeneous strain imposed amounts to, say, OA, the stress–strain curve obtained will not coincide with the 'basic' curve, but will show a stress level difference of A′A″. This reflects the amount of redundant shearing. A shift of the curve to the right, until the two coincide (the Hill–Tupper principle), gives the true value of the strain imposed, by including the redundant component AB.

In hot processes, the level of stressing is of less importance, but the presence of shear affects recrystallisation, gives rise to internal faults, and further weakens the structure which may already be affected by the presence of inclusions, etc. These observations are summarised schematically in *Figure 4.17*.

To assess *a priori* the suitability of the given material for processing in a specific forming system, the material properties must be either known or will have to be determined. Standard, tensile, compressive and torsional laboratory tests may have to be conducted, but these, in turn, must be chosen carefully since not every test will necessarily be suitable. For example, rotary operations proceed mainly in torsion and, consequently, material properties established in the torsional mode will give a better idea of material response than, say, tensile or compressive testing. The selection of the correct type of test may be based on the information provided in *Table 4.3*.

4.3 An outline of the theory of plasticity

4.3.1 Basic concepts

Once the initial yield point of the material has been exceeded and the load is allowed to continue to increase, plastic flow is initiated. In strain-hardening materials this produces an ever-increasing level of the current flow stress and, also, an internal consolidation of the material. The latter physical fact provides the basic law of plasticity, i.e. constancy of volume. If the three, mutually perpendicular strains are ε_x, ε_y, ε_z, the law is given in the form:

$$\varepsilon_x + \varepsilon_y + \varepsilon_z = 0 \qquad (4.1a)$$

The characteristic behaviour of the material in the plastic regime can differ from its behaviour in the elastic range in that a reversal of strain (in direction) is possible and the strain path need not be straight since a rotation of the successive planes of the loaded specimen can take place. To account fully for this possible pattern, it is *insufficient* to consider the initial and final stages of the deformation only, but it is necessary to follow the deformation of an element through the zone of

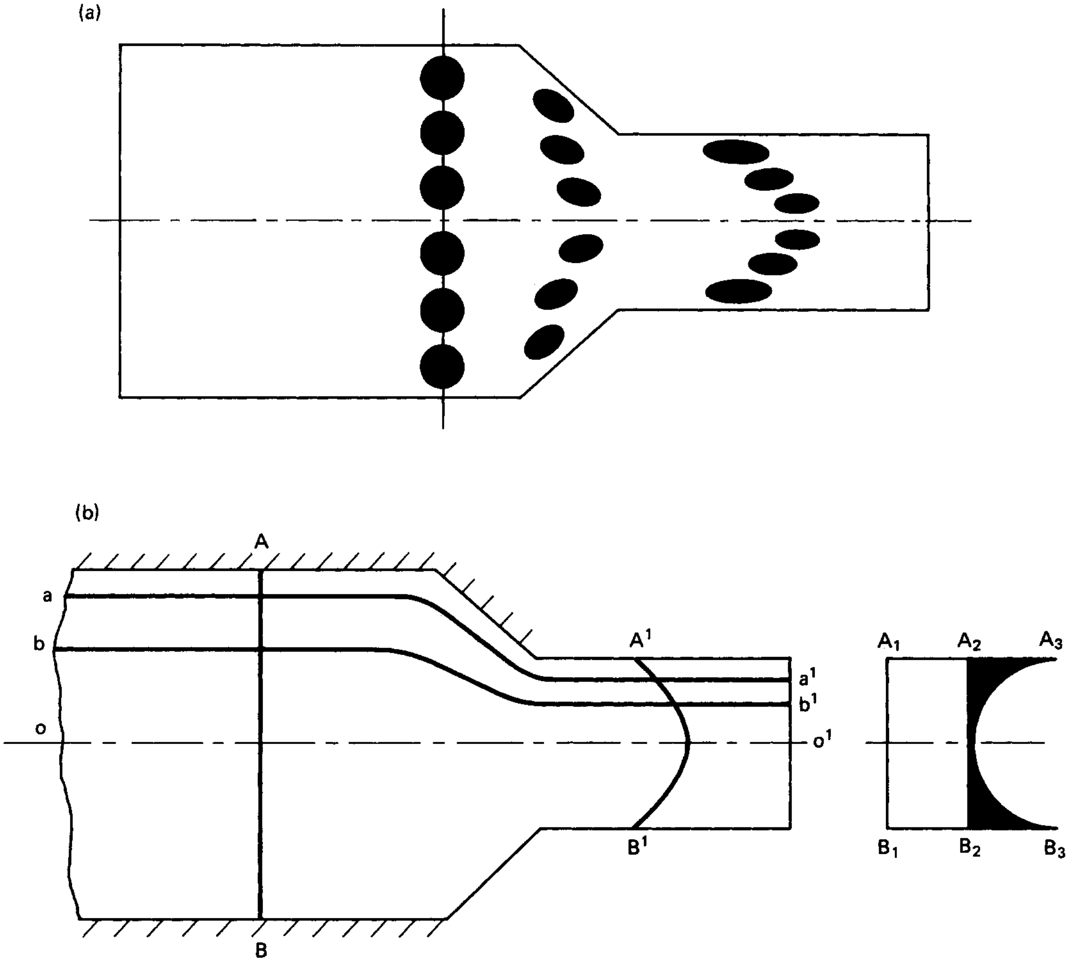


Figure 4.15 The pattern of shearing strains expected in bar drawing or extrusion

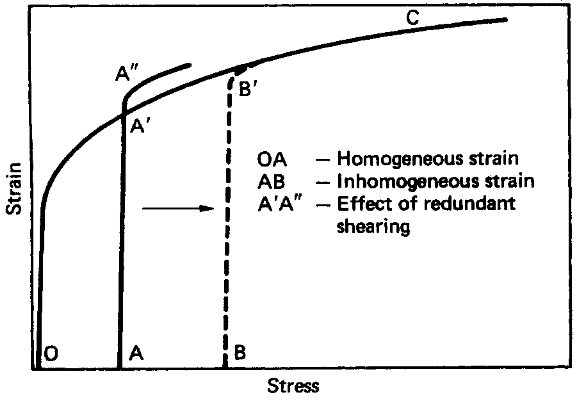


Figure 4.16 The effect of redundancy on the flow stress of a strain-hardening material

deformation in an incremental manner. In other words, analysis of the pattern of flow has to be based on incremental and *not* finite changes.

Equation (4.1a) is therefore often quoted in the form:

$$d\varepsilon_x + d\varepsilon_y + d\varepsilon_z = 0 \tag{4.1b}$$

where the strain increments have to be defined, in each individual case, in terms of the variable dimensions of the worked specimen. For instance, in the rod or wire drawing, shown in *Figure 4.11*, the radial strain increment ($d\varepsilon_y$) would reflect the change in the radius (R) varying through the zone of deformation, the axial increment ($d\varepsilon_x$), the variation in the cross-sectional area (A), the circumferential increment ($d\varepsilon_z$) and the change in the circumference (S). Hence

$$d\varepsilon_x = \frac{dA}{A};\ d\varepsilon_y = \frac{dR}{R};\ \varepsilon_z = \frac{dS}{S} \tag{4.2a}$$

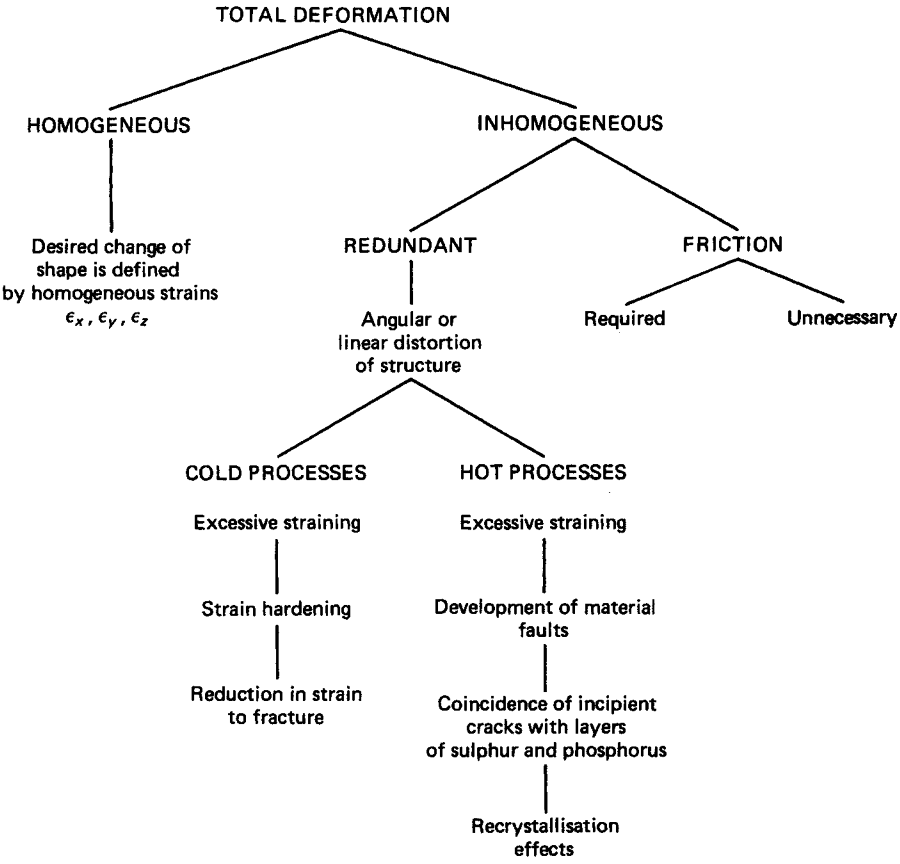


Figure 4.17 Components of total deformation

But, $A = \pi R^2$, and $dA = 2\pi R dR$, and $S = \pi R$ and so $dS = \pi dR$. On substituting these values into equation (4.2a), and accounting for the constancy of volume (equation (4.1a)), the strain increments become:

$$d\varepsilon_x = \frac{2dR}{R};\ d\varepsilon_y = -\frac{dR}{R};\ d\varepsilon_z = -\frac{dR}{R} \tag{4.2b}$$

This approach is adopted in all metal-forming processes, but naturally the expressions obtained are generally more complex. *Table 4.4* gives the values of strain increments for a selected number of more common processes.

On integrating individual expressions of the type given by equation (4.2a) between the initial and final values of the considered sections, finite values of strains are obtained in natural-logarithmic form. Strains so evaluated are known as 'natural', 'logarithmic' or 'true' since, unlike engineering percentage strains, they define the actual deformation in any given section.

The true strains correspond to specific loading conditions which result in the presence of stresses. In the plastic regime, changes in the cross-sectional area, material thickness, etc., are so substantial that the concept of 'nominal' stress, based on the original dimensions becomes untenable and its use introduces considerable numerical error. All the stresses must be calculated on the basis of the current, actual dimensions.

Although the onset of plastic flow is usually the result of the interaction of different stress fields, it is quite possible to visualise this situation as arising from the application of a single force of sufficient magnitude to cause the material to flow. To have the desired physical effect, the stress associated with this hypothetical force must assume the same value as the yield or flow stress Y. This particular stress, known as the 'generalised stress' or 'equivalent stress', is often denoted by the symbol $\bar{\sigma}$. It is used in calculations concerned with the plastic regime of flow and, in terms of the principal stresses, is given by:

$$\bar{\sigma} = \frac{\sqrt{2}}{2}[(\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_3 - \sigma_1)^2]^{1/2} \tag{4.3}$$

Similarly, in order to combine the effect of the individual strains—to be able to consider the total deformation of the specimen—the concept of the generalised or equivalent strain is introduced. In terms of homogeneous strains only, the homogeneous strain increment is given by:

$$d\varepsilon_H = \frac{\sqrt{2}}{3}[(d\varepsilon_x - d\varepsilon_y)^2 + (d\varepsilon_y - d\varepsilon_z)^2 + (d\varepsilon_z - d\varepsilon_x)^2]^{1/2} \tag{4.4a}$$

Table 4.4 Strain and strain-rate relationships

Process	*Homogeneous strain increments*	*Generalised homogeneous strain*	ε_m
Rotary piercing	$d\varepsilon_x = -\dfrac{D_0 - 2h}{h(D_0 - h)}\,dh$ $d\varepsilon_y = \dfrac{dh}{h}$ $d\varepsilon_z = -\dfrac{dh}{d_0 - h}$ $d\varepsilon_H = \dfrac{2\sqrt{3}}{3}\left[\dfrac{3h^2 - D_0h + D_0{}^2}{h^2(D_0 - h)^2}\right]^{1/2} dh$	$\varepsilon_H = \ln\left[\left(1 - \dfrac{8BC}{(C - 3B)(C - B)}\right)^{1/\sqrt{3}}\left(\dfrac{D_0}{C - D_0 - \sqrt{3}B}\right)\right]^2$ $B = (2h_1 - D_0);\ C = (3h_1^2 - 3D_0h_1 + D_0^2)^{1/2} + D_0$	$\dot{\varepsilon}_m = \dfrac{aV_0}{2\sqrt{3}[-b + (G + 4ah)^{1/2}]}\dfrac{2B}{(1 - B^2)} + \ln\dfrac{(1 + B)}{(1 - B)}$ $h = ax^2 + bx + c;\ B = \dfrac{2h - D_0}{(3h^2 - 3D_0h + D_0^2)^{1/2}};$ $G = b^2 - 4ac$
Secondary rotary piercing	$d + h = \text{constant}$ $d\varepsilon_x = -\dfrac{dh}{h}$ $d\varepsilon_y = \dfrac{dh}{h}$ $d\varepsilon_z = 0$	$\varepsilon_H = \dfrac{2\sqrt{3}}{3}\ln\left(\dfrac{h_0}{h_1}\right)$	—
Assel elongation	$d\varepsilon_y = -\dfrac{dh}{h}$	$\varepsilon_H = \ln\left\{\dfrac{(1 + B_0)(1 - B_n)}{(1 - B_0)(1 + B_n)}\right.$	$\dot{\varepsilon}_m = (\beta \text{ or } \gamma \text{ or } \delta)\left\{\dfrac{2(B_n - B_0)(B_0B_n + 1)}{(1 - B_0^2)(1 - B_n^2)} + \ln\dfrac{(1 + B_n)(1 - B_0)}{(1 + B_0)(1 - B_n)}\right\}V_0$
Tube extrusion Mandrel and plug tube drawing	$d\varepsilon_x = \dfrac{d + 2h}{(d + h)h}\,dh$ $d\varepsilon_z = -\dfrac{dh}{d + h}$ $d\varepsilon_H = \dfrac{2}{\sqrt{3}}\left[\dfrac{3h^2 + 3hd + d^2}{(d + h)^2h^2}\right]^{1/2} dh$	$\left.\left[\dfrac{(1 - \frac{2}{\sqrt{3}}B_0)(1 + \frac{2}{\sqrt{3}}B_n)}{(1 + \frac{2}{\sqrt{3}}B_0)(1 - \frac{2}{\sqrt{3}}B_n)}\right]^{2/\sqrt{3}}\right\}$ $B = \dfrac{\sqrt{3}(2h + d)}{2(3h^2 + 3hd + d^2)^{1/2}}$	$\beta = \dfrac{D_0^2 - d^2}{4\sqrt{3}\,d^2(x_n - x_0)};\ \gamma = \dfrac{a(D_0^2 - d^2)}{2\sqrt{3}\,d^2[-b + (G + 4ah_n)^{1/2}]};$ $\delta = \dfrac{(D_0 - h_0)h_0 \tan\alpha}{\sqrt{3}\,d^2(h_0 - h)};\ B = \dfrac{2h + d}{(3h^2 + 3hd + d^2)^{1/2}}$ a, b and G as in rotary piercing
Tube sinking	$d\varepsilon_x = \dfrac{dD}{D - h_0}$ $d\varepsilon_y = \dfrac{dh}{h} = 0$ $d\varepsilon_z = -\dfrac{dD}{D - h_0}$ $d\varepsilon_H = \dfrac{2}{\sqrt{3}}\left(\dfrac{dD}{D - h_0}\right)$	$\varepsilon_H = \dfrac{2\sqrt{3}}{3}\ln\left(\dfrac{D - h_0}{D_0 - h_0}\right)$	$\dot{\varepsilon}_m = \dfrac{4V_0 \tan\alpha}{3(D - h_0)}$
Rod extrusion	—	—	
Bar and wire drawing	—	—	$\dot{\varepsilon}_m = \dfrac{2(D_0 + D)V_0 \tan\alpha}{D^2}$

In many actual operations in which rotation of the axes of strains is either absent or negligibly small (such as in wire and tube drawing and extrusion, or strip rolling) and in which the strain path is clearly defined, equation (4.4a) can be integrated directly to give:

$$\varepsilon_H = \sqrt{\frac{2}{3}}(\varepsilon_x^2 + \varepsilon_y^2 + \varepsilon_z^2)^{1/2} \tag{4.4b}$$

The final, generalised strain, including the redundant strain component, is given by:

$$\varepsilon_T = \sqrt{\frac{2}{3}}[\varepsilon_x^2 + \varepsilon_y^2 + \varepsilon_z^2 + \tfrac{1}{2}(\varepsilon_l^2 + \varepsilon_c^2 + \varepsilon_r^2)]^{1/2} \tag{4.4c}$$

where ε_l, ε_r and ε_c are the inhomogeneous longitudinal, radial and circumferential strains, respectively. Since the shears produce only internal distortion of the structure and do not influence dimensional changes, the principle of constancy of volume does not apply to them and they can thus appear singly, in pairs, or all three may be present with the mathemtical linkage missing. These strains must be assessed experimentally in each individual case. The values of the generalised, homogeneous strains for a number of the forming processes, are given in *Table 4.4.*

The relationship between the generalised, or flow stress and true, generalised strain is known as the 'constitutive equation' of the material and is characteristic of the material behaviour in specified working conditions. Equations of this type are determined experimentally, using standard laboratory test techniques (*Table 4.3*). Generally, the engineering alloys used in metal forming fall into five distinct groups according to their properties. In terms of their constitutive relations, they can be classified as:

(1) perfectly elastic (no plastic ductility);
(2) elastic, but linearly strain hardening in the plastic range (approximation to, say, ferrous alloys);
(3) elastic–perfectly plastic (a critical and constant value of stress produces flow, until fracture occurs);
(4) rigid–perfectly plastic (negligible elasticity); and
(5) rigid–linear strain hardening (negligible elasticity).

These properties are illustrated in *Figure 4.18.*

The usual, empirical equations representing the properties of an actual material are of the power-law type, and the functional relationship between the flow stress Y and the total, generalised strain is either in the form

$$Y = A\varepsilon^n \tag{4.5}$$

or

$$Y = B(C + \varepsilon)^n \tag{4.6}$$

where A, B and C are constants related to the material properties, and n ($0 \leq n \leq 1$) is the exponent indicating the strain-hardening characteristics.

In the condition of increasing strain rate, the response of the material is conditioned by the properties of the forming system (both in terms of geometry and kinematics) and the influence of frictional and internal shearing parameters. The effect of the rate of strain can be represented either by the instantaneous or the mean strain rate. Using the concept of generalised homogeneous strain and considering the time in which it takes place, the instantaneous rate ($\dot{\varepsilon}_H$) is given by

$$\dot{\varepsilon}_H = \frac{d\varepsilon_H}{dt} \tag{4.7}$$

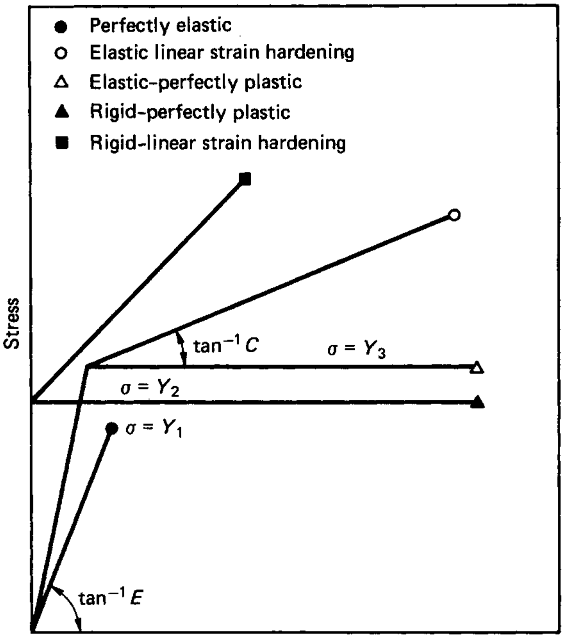


Figure 4.18 Idealised representation of basic stress–strain relationships

and the mean strain rate $\dot{\varepsilon}_m$ is given by

$$\dot{\varepsilon}_m = \frac{1}{t}\int_0^t \dot{\varepsilon}_H dt \tag{4.8}$$

Since the strain rate is defined in the manner of equation (4.7), it does not reflect the variation in the velocity of the strained elements as they proceed along the pass. Furthermore, the mean strain rate (equation (4.8)) refers solely to the material element under consideration, but not to the volume of the material undergoing instantaneous deformation.

In metal forming, it may be more important to consider the strain rate prevailing in the whole zone of deformation, rather than just locally, since this will account for the variation in velocity. This objective is attained if the mean strain rate is referred to the physical boundaries of the pass, rather than to the time elapsed (t). In this case,

$$\dot{\varepsilon}_m = \frac{1}{x}\int_0^x \dot{\varepsilon}_H dt \tag{4.9}$$

The values of the latter strain rate are quoted in *Table 4.4.*

4.3.2 Yielding and criteria of yielding

The elastic 'failure' or the onset of plasticity are determined for a given material by using an appropriate criterion of yielding which defines the value of the critical flow stress as a function of the total strain.

The simpler of the two criteria, and that most commonly used, is the maximum shear stress (or the Tresca criterion). This ignores the effect of the third principal stress and is concerned solely with the influence that the maximum possible stress difference has on the stability of the molecular planes.

The Tresca criterion is defined as:

$$|\sigma_3 - \sigma_1| = Y(\varepsilon) = k(\varepsilon) \tag{4.10}$$

where k is the yield stress in shear.

The more general Huber–Mises criterion is based on the critical value of the shear strain energy that can be absorbed by the material within the elastic range of response, and is given by:

$$(\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_3 - \sigma_1)^2 = Y'(\varepsilon) = k'(\varepsilon) \tag{4.11}$$

The use of a mathematical function will account for the changes in the level of the yield stress, but the function itself may be difficult to handle in actual calculations. In recognition of this difficulty, a concept of the *mean* yield stress, operating over the working zone of the pass, can be used.

A mean yield stress Y_m is defined as:

$$Y_m = \frac{1}{\varepsilon}\int_0^{\varepsilon} Y(\varepsilon)d\varepsilon \tag{4.12}$$

where ε is the total, generalised strain of equations (4.5) and (4.6). Alternatively, the value can be obtained by employing the Hill and Tupper approach and utilising the information available in a plot of the type represented in *Figure 4.16*. When the translation of the actual stress–strain curve is effected, the values of the mean flow stresses, in addition to those of the component strains, can be determined by reference to the areas enclosed under the respective curves. Thus, in the absence of any redundant effect, the mean flow stress would be Y_H which is obtained by measuring the area OA'A and dividing it by ε_H (OA). In the presence of redundancy, the mean flow stress (Y_T) is defined as the area OB'B/ε_T, and the proportion of the stress induced by redundancy Y_R = (area AA'B'B)/ε_R. Where ε_R = AB.

4.3.3 Redundancy and geometry factors

The level of redundancy in an operation has to be assessed in order to provide a means of determining the efficiency of the process. This, as already indicated, depends on the geometry of the pass and the properties of the material.

For most metal-forming processes, when the total strain is plotted against the homogeneous deformation, the former increases at a higher rate than would be expected for an efficient forming process. The relationship is practically linear, but the ratio of the two quantities, for any given value of the homogeneous deformation, can be high. This type of relationship leads to the definition of the 'strain redundancy factor' ϕ:

$$\varepsilon_T = \phi\varepsilon_H \tag{4.13}$$

The factor is clearly a function of both the strain and the geometry of the pass and so can be defined further as:

$$\phi = \phi(\varepsilon,\Delta) \tag{4.14}$$

where Δ is the geometry factor.

Therefore the factor can be represented by both a linear function (a reasonable approximation for most bulk metal-forming processes) and, by a modified expression that involves the geometry factor. Hence,

$$\varepsilon_T = K_1\varepsilon + K_2 \sin\alpha$$

or

$$\phi = K_1 + K_2\frac{\sin\alpha}{\varepsilon} \tag{4.15}$$

and

$$\phi = C_1 + C_2\Delta \tag{4.16}$$

where K_1, K_2, C_1 and C_2 are constants depending on the material and the frictional conditions.

Equation (4.15) is appropriate for drawing and extrusion. Similar expressions can be obtained for rotary and longitudinal rolling processes but, of necessity, can be, and often are, more complex.

The equivalence of the expressions given by equations (4.15) and (4.16) must be proved in each individual case and will depend on the somewhat arbitrary definition of the geometry factor.

The case of drawing represented in *Figure 4.11*, can again be taken as an example. On substituting equation (4.2b) into (4.4a), and integrating the latter within the limits of the initial and final diameters D_o and d, the generalised homogeneous strain is obtained as:

$$\varepsilon_H = 2\ln(D_o/d) \tag{4.17}$$

The geometry factor for a conical die is often defined as:

$$\Delta = \frac{D_o + d}{D_o - d}\sin\alpha \tag{4.18}$$

where α is the die semi-angle. Consequently, equation (4.16) in conjunction with equation (4.17) gives:

$$\phi = C_1 + C_2\left(\frac{e^{\varepsilon/2} + 1}{e^{\varepsilon/2} - 1}\right)\sin\alpha$$

$$= C_1 + C_2\left(\frac{1}{\tan h\,(\varepsilon/4)}\right)\sin\alpha$$

For small die angles and normal bar-drawing practice, expansion of this function gives

$$\tan h(\varepsilon/4) \simeq \varepsilon/4$$

and so

$$\phi = C_1 + 4C_2\frac{\sin\alpha}{\varepsilon_H} \tag{4.19}$$

It is clear that equations (4.15) and (4.19) are equivalent, with $K_1 = C_1$, $K_2 = 4C_2$, and $\Delta = (\sin\alpha)/\varepsilon_H$.

Similar arguments are applied when considering other metal-forming processes.

The strain redundancy factor is the more often used in practice since it is based on measurable quantities. However, occasionally a *stress* redundancy factor Φ is introduced. This is defined as:

$$\sigma = \Phi\sigma_H \tag{4.20}$$

where σ and σ_H represent the operational stress levels referred to the total and homogeneous force requirements, respectively. Since stresses can only be calculated, their numerical assessment depends on the type of mathematical model or theory selected and will, therefore, be more arbitrary than strain values. The two factors are not numerically equivalent unless, of course, there is no measurable redundancy in the system in which case their value amounts to unity.

Definitions of geometry factors are somewhat arbitrary but, even so, they take into account the influence that the individual process parameters are likely to have on the conduct of the operation and the quality of the processed product. A selection of the expressions defining the individual geometry factors is given in *Table 4.5*.

Table 4.5 Geometry factors in basic processes

Process	Δ
Rotary piercing*	$\frac{1}{2}\left(\frac{D_0/2 + h}{D_0/2 - h}\right) \cot \alpha$
Wire and bar drawing and extrusion	$\frac{D_0 + d}{D_0 - d} \sin \alpha$; or $\frac{1}{2}\left(\frac{D_0 + d}{D_0 - d}\right)\left(\frac{1 - \cos \alpha}{\sin \alpha}\right)$
Tube drawing	
Sinking	$\frac{4h_0[(D_0 + d) - h_0]}{(D_0 + d)(D_0 - d)} \sin \alpha$
Plug or mandrel	$\frac{1}{2}\left(\frac{h_0 + h}{h_0 - h}\right) \sin \alpha$
Tube extrusion†	$\frac{1}{2}\left(\frac{h_0 + h}{h_0 - h}\right)\left(1 - \frac{1}{b}\right) \sin \alpha_e$ $\sin \alpha_e = \left[1 - \frac{1}{24a^2 \sin^2 \alpha}\right] \sin \alpha$

* α is the feed angle as in *Figure 4.7*.
† $a = \rho(h_0 - h)$; $b = (D_0 + d)/(h_0 + h)$; ρ is the radius of die curvature.

4.3.4 Stress–strain relationships

Unlike the elastic regime, the relationship between stress and strain is not necessarily unique once the yield point of the material has been exceeded. The two parameters are firmly linked at any instant, but the nature of this can change and does not therefore constitute a characteristic material property equivalent to, say, the Young's, rigidity or bulk moduli which define the conditions in the elastic range.

When the effect of elastic strain is negligible, as in the forming of metals, the general relationship between the two parameters is governed by the Levy–Mises rule of low, defined as:

$$\begin{aligned} d\varepsilon_1{}^p &= \lambda[\sigma_1 - \nu(\sigma_2 + \sigma_3)] \\ d\varepsilon_2{}^p &= \lambda[\sigma_2 - \nu(\sigma_3 + \sigma_1)] \\ d\varepsilon_3{}^p &= \lambda[\sigma_3 - \nu(\sigma_1 + \sigma_2)] \end{aligned} \tag{4.21a}$$

where the superscript 'p' indicates purely plastic conditions, subscripts 1, 2 and 3 refer to the mutually perpendicular, principal directions, and $\lambda = d\bar{\varepsilon}/\sigma$ is the *instantaneous* constant. This brings in both the differential nature of strain and the generalised stress and strain—as defined by the appropriate constitutive equation. Note that, again unlike the elastic range, an increment of strain is proportional to finite stress.

Since the material is incompressible, Poisson's ratio ν has a single value of 0.5, irrespective of the metal considered.

By eliminating λ from equation (4.21a), the following standard expression is obtained:

$$\frac{d\varepsilon_1{}^p}{\sigma_1 - \frac{1}{2}(\sigma_2 + \sigma_3)} = \frac{d\varepsilon_2{}^p}{\sigma_2 - \frac{1}{2}(\sigma_3 + \sigma_1)} = \frac{d\varepsilon_3{}^p}{\sigma_3 - \frac{1}{2}(\sigma_1 + \sigma_2)} = \lambda$$

$$= \frac{d\bar{\varepsilon}}{\sigma} \tag{4.21b}$$

The relationship between and interdependence of the stresses are defined by the selected criterion of yielding (equations (4.10) and (4.11)) or by a suitable mathematical model of the operation.

4.3.5 Mathematical modelling of metal-forming processes

A number of analytical methods have been developed for assessing the various parameters of metal-forming processes, showing a varying degree of accuracy and complexity. All these methods aim ultimately at relating the required deformation of the specimen to the necessary external deforming agent that must be supplied by the plant.

At present, five basic methods of analysis are usually employed. In chronological order of development these are:

(1) the equilibrium technique,
(2) the slip-line approach,
(3) the upper-bound method,
(4) the visioplasticity technique, and
(5) the finite-element method.

The equilibrium method is based on the stress field that develops in the processed material, whereas the other four are associated with phenomena related to metal flow. Irrespective of the method of analysis adopted, the basic characteristics of the plastic state of the material (i.e. the condition of constancy of volume, the yield criterion and the functional yield, flow or constitutive stress equation) will have to be assumed.

4.3.5.1 The equilibrium technique

This approach, sometimes referred to as the 'stress' or 'slab' technique, consists basically of isolating a representative volume element in the body of the material undergoing plastic deformation, and observing its behaviour as it moves along the working zone of the pass. Since in a normal physical situation the element does not cease to form an integral part of the whole of the material body, it must clearly remain in the state of force equilibrium throughout its period of deformation. The behaviour of the element, reflecting that of the whole of the worked specimen, can therefore be analysed by considering the equilibrium of forces acting on it at any instant of deformation.

A volume element of a material undergoing axisymmetric-type deformation is shown in *Figure 4.19*. The equation of equilibrium of forces in one direction is

$$\frac{\sigma_r - \sigma_\theta}{r} + \frac{\partial \sigma_r}{\partial r} - \frac{2\tau}{t} = 0 \tag{4.22}$$

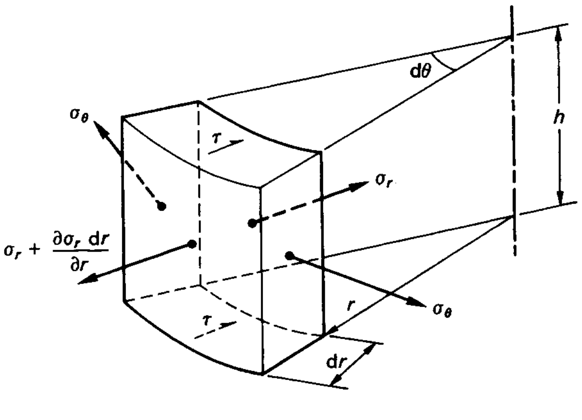


Figure 4.19 An element of material in axisymmetric equilibrium analysis

Basically, the equilibrium method requires that conditions of force equilibrium in three directions be established, i.e. $\Sigma F_x = \Sigma F_y = \Sigma F_z = 0$, or in, say, cylindrical co-ordinates that $\Sigma F_r = \Sigma F_\theta = \Sigma F_z = 0$. The analysis usually leads to the formulation of one or more differential equations involving the stress system produced in the body of the material. The traction or friction effect, originating at the tool–work piece interface, is normally allowed for by introducing a friction term incorporating a value of the coefficient of friction. The relationship between the stresses produced is established by means of a suitable criterion of yielding, ultimately related to the tensile/compressive or shear yield stress of the material. Finite values of the operative stresses are obtained by integrating differential equations between the conditions obtaining at the entry to and exit from the working zone of the pass or, if this is complex in nature, by splitting the pass into a number of interconnected zones and summing the effects in the individual zones.

The method is only approximate and underestimates the magnitudes of the forces involved by ignoring the effects of redundancy and of the pattern of flow and, therefore, by taking an oversimplified view of the mechanics of the process in question. Since, however, the analysis involved is usually reasonably straightforward, the technique is widely used whenever approximate solutions are acceptable. It is, for instance particularly suitable for wire and tube drawing, hot and cold rolling of strip and sheet, and also gives an acceptable degree of approximation for rotary piercing and elongation of seamless tubing.

If the effects of redundancy in any given situation are either known or assessable, the accuracy of force predictions given by the equilibrium approach can be improved. In addition, a separation of the homogenous effect from the inhomogeneous one can be carried out, and the respective components of work done in changing the shape and in distorting the structure can be determined. A deforming 'efficiency' of the process can thus be established.

To improve the accuracy of calculations a modification of the value of the operative stress resulting from the integration of the equation of equilibrium can be made, either by introducing the stress redundancy factor Φ (equation (4.20)) or by modifying the value of the yield stress. In the latter case, equation (4.13) (by giving the value of the total strain) will either provide the limits of integration of the yield-stress function or will give the mean operative value of the stress Y_m. Either of these can be introduced into the analysis when a criterion of yielding has been selected.

An estimate of the deforming 'efficiency' of the operation can be made in the following manner. The total operative stress, and therefore by implication the magnitude of the total deforming agent, can be calculated from the final expression given by the analysis. If, then, the friction term is omitted from the equation of equilibrium, the equation will define the stress of total deformation only. To separate the effect of the homogeneous deformation from that of the redundant one, the equation of equilibrium for frictionless forming is used once again. This time, however, the 'lower' value of the yield is introduced, i.e. the value that corresponds to the magnitude of the homogeneous strain only and is obtained from equation (4.12). The two expressions give the magnitudes of the loads necessary to produce the total homogeneous deformation. The numerical difference between the two will, of course, give the load of the redundant deformation. Work done or energy dissipated in effecting these physical changes can, therefore, be computed.

4.3.5.2 The slip-line analysis

The technique used here is based on the physical observation that plastic flow occurs predominantly as a result of microscopic slip (on an atomic scale) along crystallographic planes. A polycrystalline metal will normally contain a high number of planes of close-packed atoms, along which slip is likely to occur, a sufficient number of these being orientated in such a way that, on a macroscopic scale, the general direction of slip will coincide with that of the maximum shear stress. The directions of the maximum shear stresses thus become characteristic properties of the system and, when established, lead to a numerical assessment of stress distribution throughout the deforming region including, of course, the boundaries on which loads exerted by tools are acting. Calculations of such loads becomes possible.

Detailed descriptions of the theoretical bases of the technique, its scope and limitations, the practical methods of application to a variety of problems and the derivation of the appropriate equations can be found in the standard textbooks on the subject (see Bibliography). It is sufficient to note here that this particular method of analysis applies rigorously only to plane strain deformations of a rigid, perfectly plastic isotropic solid, and that, unless artificially modified, it cannot account for strain hardening, strain rate, or frictional and redundancy effects.

The method forms a useful tool in the general field of metal forming, mainly by giving some insight into the pattern of flow of the material. The basic difference between a slip-line approach and that of, say equilibrium-type solution lies, however, in the fact that a specific slip-line field has to be postulated *a priori*, and that its feasibility must then be checked against the existing experimental data. It is clear that in the general course of events the method cannot give a unique solution, and that a number of fields could be proposed in a specific case that would fit the specified system of boundary conditions equally well. In fact, a criterion of minimum work must be invoked to help select the solution that is most likely to be the probable one.

To obtain a solution, families of curvilinear or straight lines are constructed that intersect orthogonally and correspond to the directions of maximum and minimum constant shear lines. Static equilibirum, the yield condition and the pattern of flow in the plastic zone must be satisfied by the assumed network of these lines. The mutually perpendicular lines in any x–y plane are designated as α and β.

Deformation in a plane strain situation is that of pure shear and, consequently, it is normally assumed at this point that the deformation is the result of pure shear stress. Yielding will occur when the maximum shear stress reaches the appropriate value k, which is limiting for the given material. Since no linear strain can be experienced along the slip lines, an element formed between two α and β lines, say that shown in *Figure 4.20*, can only distort and will neither extend nor contract. If σ_1, σ_2 and σ_3 are principal stresses, then in plane strain $\sigma_2 = \frac{1}{2}(\sigma_1 + \sigma_3)$. This can be expressed in terms of hydrostatic compressive pressure $(-p)$ which is appropriate in most metal-forming operations as $\sigma_2 = -p$. The principal stresses are then given by:

$$\sigma_1 = -p - k;\ \sigma_2 = -p;\ \sigma_3 = -p + k \qquad (4.23)$$

The equations for the equilibrium of forces acting on the element shown in *Figure 4.20* are:

$$\frac{\partial p}{\partial \alpha} + \frac{2k}{r_\alpha} = 0 \text{ (along an } \alpha \text{ line)}$$

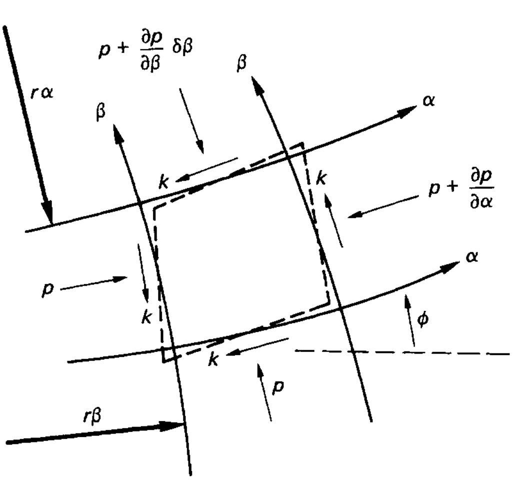


Figure 4.20 A material element between two slip lines

and

$$\frac{\partial p}{\partial \beta} - \frac{2k}{r_\beta} = 0 \text{ (along a } \beta \text{ line)}$$

Expressing the radii of curvature in terms of θ, it is found that $1/r_\alpha = \partial\theta/\partial\alpha$ and $1/r_\beta = \partial\theta/\partial\beta$, and therefore on integration

$$p + 2k\theta = C_1 \text{ (along an } \alpha \text{ line)} \quad (4.24)$$

$$p - 2k\theta = C_2 \text{ (along a } \beta \text{ line)} \quad (4.25)$$

For a hardening material, the yield stress in shear k remains constant along a slip line, but the pressure p may vary. The stress configuration can be determined at any point on the deforming material if the magnitude of p and the direction of k are known. The latter is defined by the direction of the considered slip line, whereas the former can be estimated by considering the changes in p produced by an angular rotation of the slip line between two points in the field. The actual finite value of p can be established from knowledge of the appropriate boundary conditions.

As already indicated, the validity of a slip-line field which satisfies the stress boundary conditions must be verified by means of continuity equations. In other words, a check must be made on whether a velocity field that satisfies equilibrium conditions actually exists.

In the absence of any linear changes in the dimensions of an element along a slip line, there can be no change in the direction of shear between two points on a slip line; the velocity component perpendicular to the line will vary from point to point. Denoting the velocity component parallel to the α line as u, and that parallel to the β line as v (*Figure 4.21*) it is found that the following equations hold:

$$du - v d\theta = 0 \text{ (along an } \alpha \text{ line)} \quad (4.26)$$

$$dv + u d\theta = 0 \text{ (along a } \beta \text{ line)} \quad (4.27)$$

If a material point within the deforming body suffers a change in the magnitude of its absolute velocity on crossing a slip line, the conditions of continuity demand that the components of velocity normal to the slip line be the same. However, no such stipulation is made with regard to the tangential velocity components, i.e. velocity discontinuities are possible in directions tangential to the slip lines. These conditions, used in

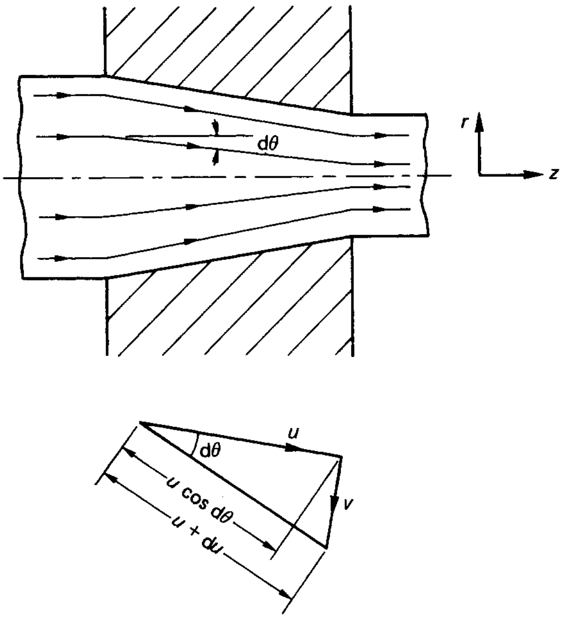


Figure 4.21 Diagrammatic representation of the velocity field of a particle in tube extrusion

conjunction with equations (4.26) and (4.27), provide the required means of testing the feasibility of the proposed slip-line field. It should, however, be remembered that a good approximation to this set of conditions can be expected only within the fully developed plastic region, and not when crossing from rigid to plastic regions of deformation where the change takes place over a finite transition zone.

To establish a slip-line field, the stress and slip boundary conditions have to be clearly specified. The following four possibilities exist:

(1) free surface,
(2) frictionless interface,
(3) Coulomb friction condition (μ is constant), and
(4) perfectly rough surface.

Free surface In the absence of a normal stress at the surface, or with $\sigma_3 = 0$, $p = k$ (equation (4.23)) if σ_1 is compressive, and $\sigma_1 = -2k$. If σ_1 is tensile, $p = -k$ and $\sigma_1 =$ k. Slip lines meet at the free surface at 45°.

Frictionless interface With the resultant shear on the surface being zero, slip lines meet again at 45°, but in the presence of a normal stress $\sigma_3 \neq 0$ and, consequently, $p \neq k$.

Correction for Coulomb friction Taking the vertical pressure at the point of contact as q, it is clear that the frictional effect μq on the surface must be balanced by stress $k \cos 2\theta$, related to the plane of maximum shear or the slip line inclined at angle θ to the interface, and consequently $k \cos 2\theta = \mu q$. The problem that immediately arises in this connection is that of knowing the value of q. Since q cannot be determined before the field has been drawn, it is necessary to adopt an iterative technique of calculation by taking the first value of q as that obtained from the consideration of a frictionless surface.

Perfectly rough surface This normally represents the condition of sticking friction or absence of interfacial movement. The frictional stress τ is equal to k, and is independent of the normal stress. The necessary slip-line configuration gives one line normal to and the other tangential to the surface.

The use of the slip-line method of analysis in the examination of problems involving large redundancies presents certain difficulties. Although the effect of friction can be incorporated and a fair degree of accuracy can be attained, the problem of strain hardening is more difficult. Hencky's equations ((4.24)–(4.27)) can be modified by the addition of another term to account for hardening, but the method is cumbersome and has so far shown promise only for machining operations. The concept of the mean yield stress is, therefore, again introduced, and an equation of the type given by (4.12) is used with substitution of $2k$ for $Y(\varepsilon)$.

Effects of redundancy are usually incorporated by correcting the value of the operative stress, determined from the field, by means of factor Φ.

The most serious limitation of the analysis, of course, is its dependence for accuracy on the conditions of plane strain. Attempts have been made to extend the technique to axisymmetric problems, and have met with some limited success in application to the extrusion of bars. A similarity between the results obtained from slip-line-theory considerations, applied to plane strain extrusion through square dies, and experimentation with axisymmetric extrusion, led Johnson (see the Bibliography) to propose a semi-empirical method of solution based on the slip-line-field approach. This method is discussed later (see Section 4.4) in the context of extrusion operations. It should be noted here that a correction for hardening was made on the assumption that the total equivalent strain is independent of hardening characteristics.

4.3.5.3 *The upper-bound method*

Within the scope of its assumptions, the slip-line-field technique produces an exact solution and, whilst this seems desirable, difficulties are often encountered in finding suitable slip-line fields for new problems. In addition, axisymmetric problems cannot be accommodated. The real power of the technique lies in the derivation of general solutions to plane strain problems.

Limit analysis has been proposed as a means of obtaining solutions to many otherwise intractable metal-forming problems. In its complete form, this approach seeks to establish two expressions for forming loads, one (the so-called 'lower bound') which is definitely an underestimate. Since it is generally sufficient to ensure that the calculated load for a forming operation is indeed sufficient to perform the process, then interest naturally centres on the upper-bound technique. If used correctly, realistic overestimates are obtained.

Upper-bound solutions involve consideration of the conditions which must be fulfilled by the strain increments in a plastically deforming medium and are not concerned with stress equilibrium. An important concept involved is that of a kinematically admissible velocity field. This is a distribution of generalised particle velocities which is kinematically compatible with itself and with the externally imposed velocities at boundaries. The total amount of work done by the externally applied forces in a metal-forming operation is used to overcome the resistance to plastic deformation and the frictional resistance to relative motion at the tool–work piece interface. These two components can be calculated from a velocity field and the upper-bound theorem that states that the kinematically admissible velocity field which minimises the work done is the actual velocity field. From such a field the actual work done can thus be derived, although in practice the actual field may never be completely attained. When formulating a solution the general procedure is to divide the deformation zone into one or more assumed zones throughout each of which the velocity is continuous. In adajcent zones a different velocity distribution may exist, whilst across these interfaces and at the tool–work piece interfaces a tangential velocity discontinuity may occur. The existence of velocity distributions in each zone implies that plastic strain and, therefore, energy must be dissipated:

$$P_1 = \int \bar{\sigma}\dot{\bar{\varepsilon}}\, \mathrm{d}V \tag{4.28}$$

where V is the volume of the deforming metal. P_1 is often referred to as the 'power of deformation'. Similarly, energy must be dissipated when material is sheared across velocity discontinuities. This term, known as the 'shear power', is given by:

$$P_2 = \int k \mid \dot{s} \mid \mathrm{d}A \tag{4.29}$$

where A is the area over which the velocity discontinuity of magnitude $\dot{s}$ occurs.

Energy dissipated due to frictional resistance along tool–work piece interfaces can be accounted for in a similar manner. A tangential velocity discontinuity $\dot{s}$ compatible with a shear stress of maximum value k for perfectly rough interfaces will occur. For Coulomb friction, a proportion mk of the shear stress can be used. Thus the friction power p_3 is given by:

$$p_3 = \int \mathrm{ck} \mid \dot{s} \mid \mathrm{d}A \tag{4.30}$$

The upper-bound theorem can be stated as:

$$P = P_1 + P_2 + P_3 \tag{4.31}$$

where P is the total power required to perform the operation.

Frequently, a characteristic parameter of the velocity field is retained as a variable when calculating the right-hand side of equation (4.31). This parameter can then be optimised to minimise the total power.

4.3.5.4 *The visioplasticity technique*

The visioplasticity technique makes use of experimental observations of flow patterns and provides a means of deriving complete solutions. However, solutions are obtained after the experimental data have been obtained. The technique cannot therefore be used to predict the likely behaviour of the processed medium, but it is very useful in a detailed analysis of the distribution of stress, strain and strain rate in any section within the deforming zone of a plastic body. Furthermore, because the analysis is based on observed experimental fact and does not, therefore, require any assumption other than the usual one regarding the stress–strain relationship, it gives an exact solution and thus can be used as a 'standard' to assess the accuracy of results obtained by other means.

The technique is based on the examination of a velocity field developed incrementally within the deforming body. In a general axisymmetric case a marking grid pattern is imprinted on the meridian plane of a cylindrical work piece undergoing a metal-forming operation and the latter is conducted by imposing on the material a unit strain at a time. A record of the change in the grid, after an increment in strain has been imparted, will give a sufficiently clear picture of the instananeous pattern of flow.

As an example of the determination of stress and strain-rate fields, consider *Figure 4.21*. If V is a known or measurable

instantaneous velocity of a particle at A, whose position is defined by co-ordinates r, θ and z, and u and v are the components of V in directions z and r, respectively, then plots of u and v against, in turn, r and z can be constructed. Strain rates at point A, and any other point in the deforming material, can then be determined from the slopes of these curves:

$$\dot{\varepsilon}_z = \frac{\partial u}{\partial z};\ \dot{\varepsilon}_r = \frac{\partial v}{\partial r};\ \dot{\gamma}_{zr} = \frac{\partial u}{\partial r} + \frac{\partial v}{\partial z} \quad (4.32)$$

If it is assumed that the material is homogenous and isotropic, the directions of the principal stresses are given by:

$$\tan 2\alpha_z = \frac{\dot{\gamma}_{zr}}{\dot{\varepsilon}_z - \dot{\varepsilon}_r}$$

where α_z is the angle between the z axis and the direction of a principal stress.

The strain-rate component $\dot{\varepsilon}_\theta$ can be found from the condition of incompressibility (constancy of volume):

$$\dot{\varepsilon}_z + \dot{\varepsilon}_r + \dot{\varepsilon}_\theta = 0$$

The effective (generalised) strain rate is given by

$$\dot{\varepsilon} = \frac{\sqrt{2}}{3}\left[(\dot{\varepsilon}_z - \dot{\varepsilon}_r)^2 + (\dot{\varepsilon}_r - \dot{\varepsilon}_\theta)^2 + (\dot{\varepsilon}_\theta - \dot{\varepsilon}_z)^2 + \frac{3}{2}\dot{\gamma}_{zr}{}^2\right]^{1/2} \quad (4.33)$$

and the generalised stress by

$$\bar{\sigma} = \frac{1}{\sqrt{2}}[(\sigma_z - \sigma_r)^2 + (\sigma_r - \sigma_\theta)^2 + (\sigma_\theta - \sigma_z)^2 + 3\tau_{zr}{}^2]^{1/2} \quad (4.34)$$

In its final form, the expression for σ_z, when the velocity field upstream is known, is given by:

$$\sigma_z = \frac{2}{3}\int_{r1}^{r}\left[\frac{\partial}{\partial r}\left(\frac{\dot{\varepsilon}_z - \dot{\varepsilon}_r}{\dot{\varepsilon}/\bar{\sigma}}\right)\left(\frac{\dot{\varepsilon}_r - \dot{\varepsilon}_\theta}{r\dot{\varepsilon}/\bar{\sigma}}\right) - \frac{1}{2}\frac{\partial}{\partial z}\left(\frac{\dot{\gamma}_{zr}}{\dot{\varepsilon}/\bar{\sigma}}\right)\right]dr - \frac{1}{3}\int_{zo}^{z}\left[\frac{\partial}{\partial r}\left(\frac{\dot{\gamma}_{rz}}{\dot{\varepsilon}/\bar{\sigma}}\right) + \frac{\dot{\gamma}_{rz}}{r\dot{\varepsilon}/\bar{\sigma}}\right]dz + \sigma_{zo} \quad (4.35)$$

where σ_{zo} is a known stress related to the conditions obtaining at the entry to the working zone of the pass.

It is clear that, because of its complexity, equation (4.35) must be integrated either numerically or graphically. In fact, the use of a computer ensures a high degree of accuracy. Equation (4.35) is, of course, easily reducible to conditions of plane strain if these become characteristic of the operation considered.

Although the visioplasticity method gives very satisfactory results, it has been applied mainly to various extrusion problems. This is due partly to the labour involved in the measurements necessary to establish various relationships, and partly to the actual integration of the final expression.

4.3.5.5 Finite-element method

With the wide availability of powerful computers, a new tool in the analysis of metal-forming processes is offered by the finite-element method (FEM). This requires that the deformation zone be divided into a number of elements that form a mesh and are connected at a specified number of nodal points.

When this has been done, and the assumed mesh pattern has influenced the outcome of the exercise, the velocity distribution for each element is assumed. The unknown velocity vectors form part of a set of simultaneous equations which are solved to give the velocity distributions and the associated stresses actually present.

Both frictional and material property elements of the working system can be accommodated, but boundary conditions—pertaining to real physical situations—must be carefully assessed *a priori*.

Considerable computing time is required and each individual process must be treated in isolation. Consequently, no general solution can be offered and details of the analysis should be obtained from specialised literature (see Bibliography).

4.3.6 Simulation of metal flow

The development of various techniques of analysis of metal flow and, in particular, of controlled tool design, has led to considerable improvements in the performance of metal-forming plant. These improvements, as already indicated, are associated with the reduction in the incidence and magnitude of inhomogenous strains in the given operation. The resulting reduction in the severity of the process leads to an increase in the output and to an improvement in the quality of the product. The success of an attempt at optimising tool design can, therefore, be assessed on the basis of a comparison between the levels of the inhomogeneous deformation.

In many instances, however, the unavailability of industrial plant for long term experimental work on the effects of varying tool profiles makes it necessary to employ model materials in laboratory conditions. Internally marked specimens (see *Figure 4.22*) can be easily made using a model material and can then be processed. The development of the flow pattern can be studied by stopping the operation before it is completed, withdrawing the partially processed product,

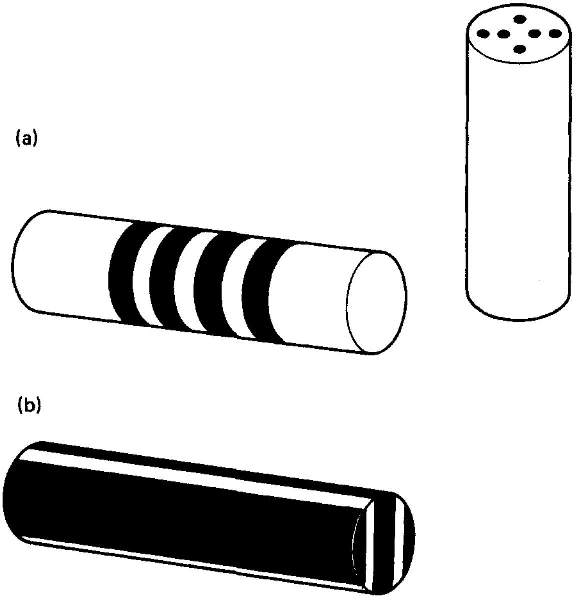


Figure 4.22 Internally marked wax billets: (a) billets for the investigation of longitudinal and circumferential shearing strains; (b) a different, more simple way of marking

and sectioning the product within the range of the working pass.

The basic requirement of a model material is that its mechanical behaviour should resemble as closely as possible that of the material which it is used to simulate. In general, no model material will comply completely with this requirement, and consequently the observed behaviour must be interpreted bearing in mind the differences existing between the model and the prototype.

If only qualitative analysis of the problem is undertaken, the discrepancies between the two materials are of less importance since they do not affect the observed pattern of flow but only the relative magnitudes of the measured quantities. In the quantitative analysis, however, it is essential to be able to 'translate' the meaning of the measurements made on the model material in terms of the properties of the prototype of the simulated material.

The magnitude of the divergence between the behaviour of a model and a prototype medium depends on the difference between the geometrical shapes of the basic true stress–strain curves, on the susceptibility of the considered materials to the effects of strain rate and/or temperature, on frictional effects, and on the work-hardening function.

For a long time, Plasticine and lead have been used as the most common model materials. It is only fairly recently that waxes and clays have been added to the list, and the possibility of using aluminium and copper has been investigated.

One of the difficulties arising from the use of non-metallic materials is their usually high degree of thixotropy. In this respect, Plasticine is less reliable since the amount of inbuilt thixotropy in the original material is unknown and the cumulative effect of further processing is impossible to determine. Waxes, however, can be standardised easily by melting and recasting.

The susceptibility of non-metallic materials to strain rates is often underestimated and considerable error may be introduced in an analysis which is based on an incorrectly selected stress–strain curve. Examples of the dependability of non-metallic materials on the conditions of testing are given in *Figures 4.23* to *4.27*.

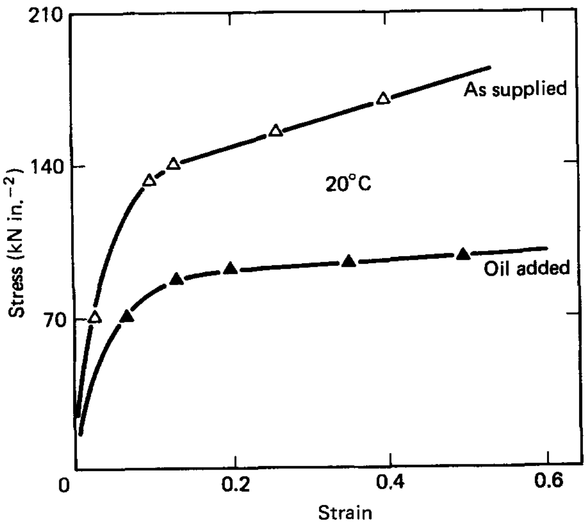


Figure 4.23 The effect of the addition of oil on the properties of Plasticine

In the case of, say, Plasticine (which is a mixture of oil and a filler), a slight change in the contents will be reflected in the different stress–strain curves (*Figure 4.23*). Equally, chemical changes produced by the dye, and the temperature level used during the test will influence the characteristic properties of the material (*Figure 4.24*).

A mixture of waxes (paraffin, ceresin, carnauba and beeswax) that has satisfactory frictional and strength properties and can simulate rotary rolling processes (which proceed in the condition of shear) can be tested in torsion. This will establish the basic properties of the mixture. In turn, these properties can be converted to the effective tensile or compressive stress, and the stress–strain curves can thus be obtained. Matching of such curves with those of the prototype or 'real' material will provide a means of selecting the correct conditions of the model.

Figure 4.25 shows the different amounts of deformation (in shear) imposed on a number of wax specimens. Shear stress–strain curves for these specimens (obtained at different shear strain rates) are shown in *Figure 4.26*. Finally, these relationships converted to the effective, tensile stress–strain quantities are given in *Figure 4.27*.

Although the incidence and pattern of the redundant deformation is independent of the material, its magnitude does depend on the constitutive relationships. Therefore the difference in response of the various materials to strain rate, temperature and frictional effects will be responsible for either a magnification of or a reduction in the effect of the shearing strain present in a given operation. A comparison between the stress–strain curves of a number of materials (*Figure 4.28*) indicates that, in general, this effect, as shown by either hot or cold wax, is likely to be exaggerated when compared with the actual effects present in the prototype. To achieve a realistic interpretation of results when using a model material which is as near as possible to the prototype in its characteristics, its stress–strain curve is matched with that of the prototype in terms of strain rate.

A convenient method of converting measurements made on the model material to the prototype has been developed and can be explained with reference to *Figure 4.29*. In this example, wax is regarded as the model (M) and lead as the prototype (P), since their respective stress–strain curves show a sufficient degree of geometrical difference to illustrate the approach. OE is the value of homogeneous strain ε_H and Y_{HP} and Y_{HM} are the mean yield stresses in the prototype and model, respectively. Y_{HP} = OACE/OE and Y_{HM} = OBDE/OE. The inclusion of inhomogeneous strains will alter the values of the respective generalised strains. ε_T will be different for each material. Let OF and OG be the strains in the prototype and the model, respectively. The mean yield stress will then be given by Y_{TP} = OAC′F/OF and Y_{TM} = OBD′G/OG. Provided that the stress–strain curves of the two materials are geometrically similar, the following approximate relationships will be true:

$$(Y_{TP}/Y_{HP}) \approx (Y_{TM}/Y_{HM}) \qquad (4.36)$$

Once the mean yield stress, corresponding to the total strain, has been determined for the model material, the value for the prototype can be estimated from an equation similar to equation (4.36), since Y_{HP} can be found directly from the appropriate stress–strain curve.

For the sake of both convenience and clarity, it can be assumed that the mean yield stress of a material is represented by two components: one that corresponds to the homogeneous deformation and the other that is induced by the presence of the inhomogeneous deformation. Using this assumption, it becomes possible quantitatively to convert inhomogeneous

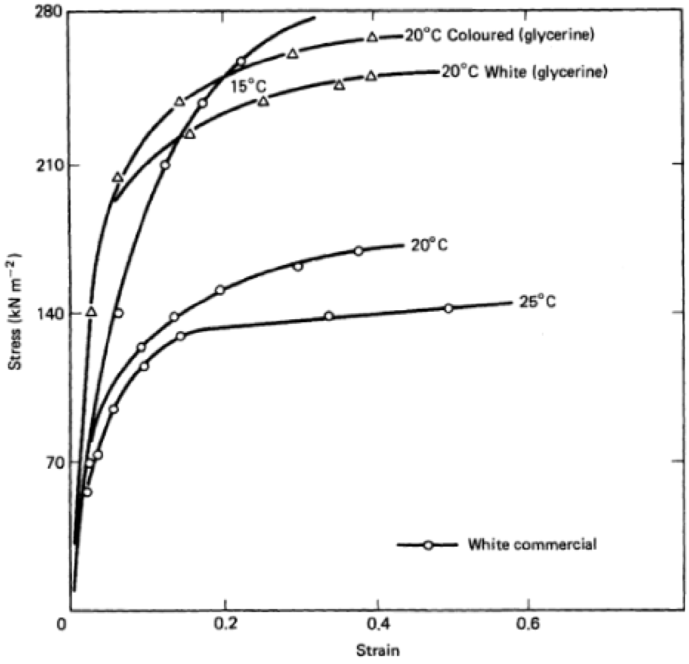


Figure 4.24 The combined effects of dying and temperature on the properties of Plasticine

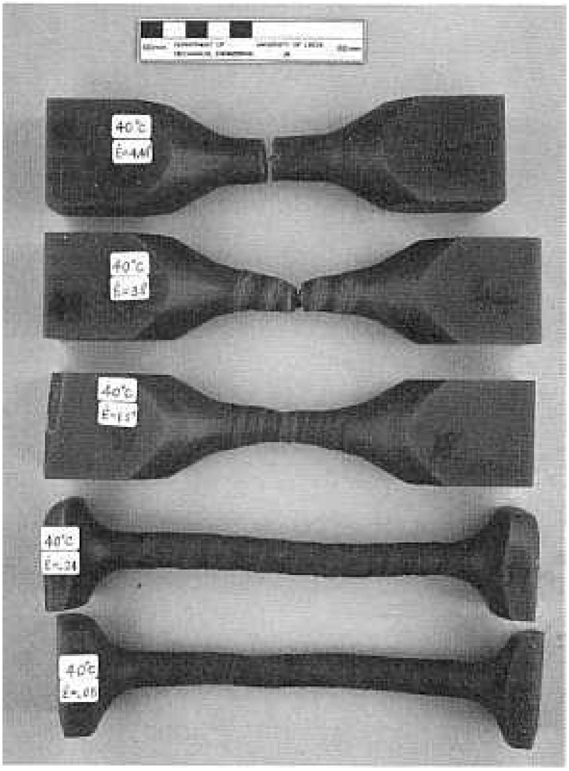


Figure 4.25 Effective strain effects in wax subjected to torsion testing at 40°C

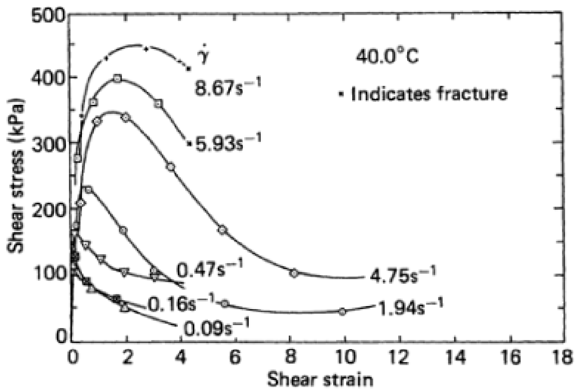


Figure 4.26 Shear stress–strain curves for the wax specimens shown in *Figure 4.25* at different shear strain rates

strains in the model to those in the prototype. Such a concept is, from the point of view of physical reality, purely artificial, but it provides a means of accounting directly for the effects of these two types of deformation.

The inhomogeneous component of the yield stress is given by Y_{IP} = ECC′F/EF and also Y_{IM} = EDD′G/EG. In general, for a material having a stress–strain curve that does not vary considerably with strain, the following relationship holds:

$$(Y_I/Y_T) = m(Y_T/Y_H) \tag{4.37}$$

where m is a correction factor determined experimentally and used if greater numerical accuracy is required. In the ordinary range of materials the value of m varies from 5% to 11%.

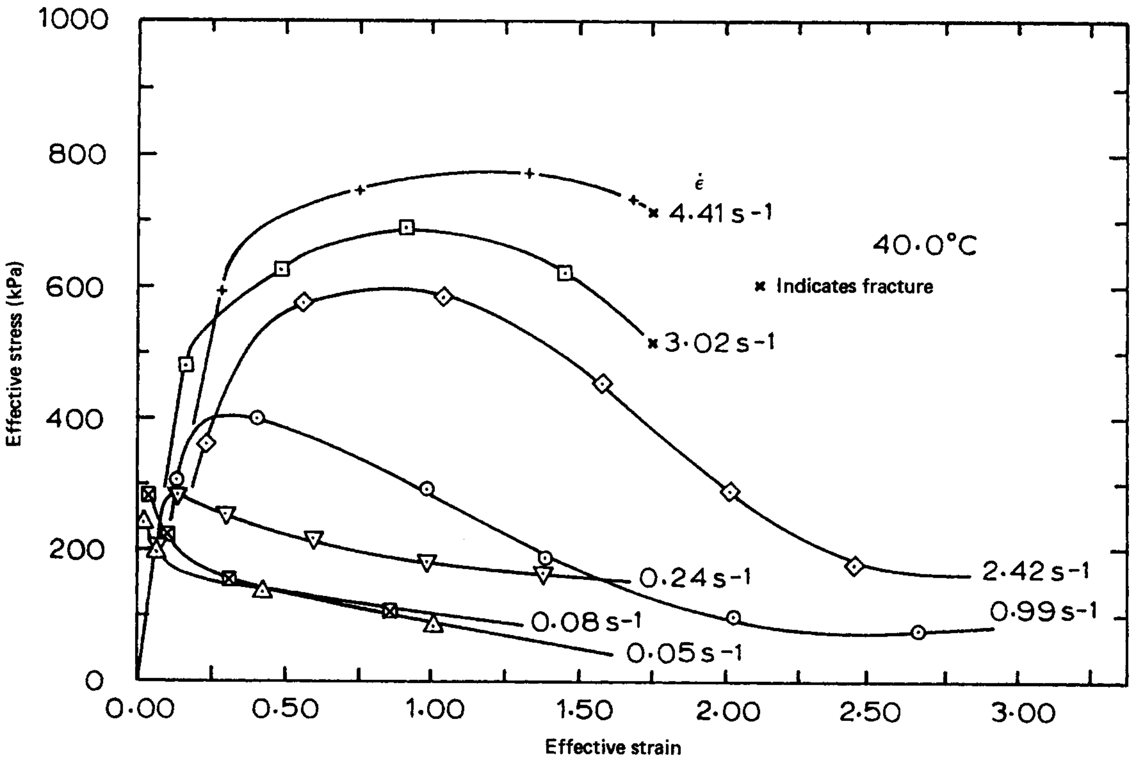


Figure 4.27 Effective stress–strain curves for the wax specimens shown in *Figure 4.25*

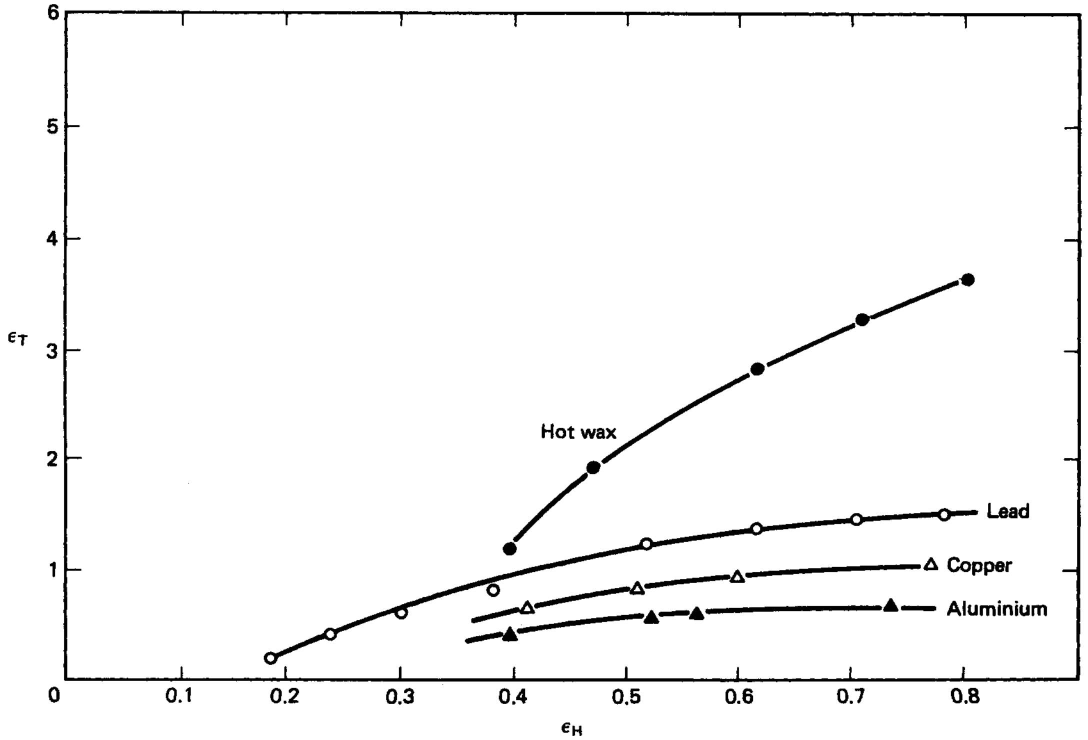


Figure 4.28 Comparison between the total and homogeneous generalised strains in wax and some metallic alloys

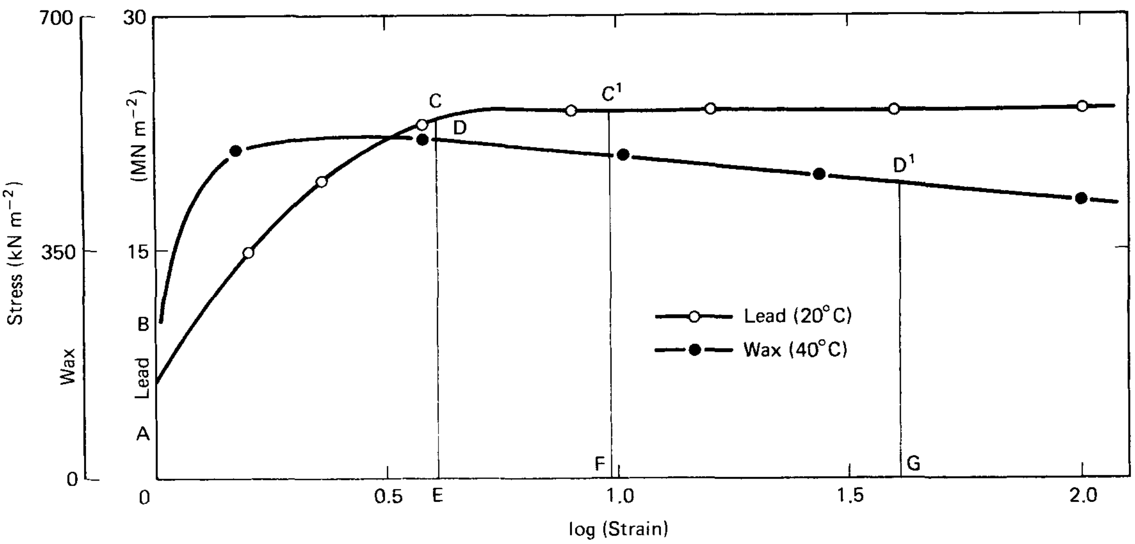


Figure 4.29 Basic and particular stress–strain curves for a strain-hardening material

Since Y_T for a material can be determined from equation (4.36), Y_I is fully defined by equation (4.37).

The convertibility of measurements made on a model material, under certain working conditions, to those likely to occur in a prototype processed under the same set of conditions can be made using equations (4.4b) and (4.4c), then

$$\varepsilon_T = \sqrt{2}[3\varepsilon_H^2/2 + \tfrac{1}{2}(\varepsilon_{xy}^2 + \varepsilon_{zx}^2 + \varepsilon_{yz}^2)]^{1/2}\sqrt{3} \qquad (4.38)$$

To determine the value of the expression in parantheses (which represents inhomogeneity) for the prototype, the value of ε_T must be evaluated. *Figure 4.16* shows that $\varepsilon_T = \varepsilon_H + \varepsilon_I$. This is numerically equivalent, in this case, to OE for both materials. The value of ε_I is, for instance, EF for the prototype and EG for the model. In general, therefore, if ε_I is measured or calculated, then if ε_H is known the total strain can be ascertained and the true value of the total inhomogeneous strain can be calculated using equation (4.38).

For any material $Y_T/Y_H = W_T/\varepsilon Y_H$ and the total work done per unit volume $W_T = W_H + W_I$. Also, in the general case, $W = Y\varepsilon$ and so

$$\varepsilon_I = \varepsilon_H(Y_H - Y_T)/(Y_T - Y_I) \qquad (4.39)$$

Again Y_I is known from equation (4.37) and, consequently, equation (4.39) can be solved. Accordingly, the quantity $(\varepsilon_{xy}^2 + \varepsilon_{yz}^2 + \varepsilon_{zx}^2)$ can be determined or, in a simple case, the actual inhomogeneous strain can be calculated.

The values of the ratios of the components of the yield stress for a few model materials are given in *Table 4.6*.

4.4 Tool design

4.4.1 Concepts of design

The problem of rational design of tools in metal-forming processes is comparatively new. Tools have been, and in most cases still are, designed on the basis of experimental selection of profiles, followed by the equally experimental 'trial and error' methods of testing their suitability. Changes in the lubrication, the range of processes materials, etc., will often call for further modifications to be made and, consequently, for the extension of the laborious and unreliable testing technique.

Table 4.6 Redundant strains in lead, aluminium and copper

ε_H	*Lead*		*Aluminium*		*Copper*		*Hot wax*
	Y_I/Y_T	Y_T/Y_H	Y_I/Y_T	Y_T/Y_H	Y_I/Y_T	Y_T/Y_H	Y_T/Y_H
0.4	1.07	1.11	1.27	1.04	1.19	1.07	1.015
0.5	1.29	1.24	1.09	1.02	0.97	0.99	1.010
0.6	1.25	1.16	1.13	1.02	0.89	0.96	0.960
0.7	1.16	1.09	1.24	1.03	0.87	0.96	0.960
0.8	1.11	1.05	1.11	1.01	0.85	0.96	0.935

In addition, the lack of appreciation of the fundamentals of the mechanics of metal flow and deformation in a given process very often results in the adoption of tool profiles which, although apparently working, impose unduly severe conditions. Tool profiles, so established, can become complicated and basically unsound.

The last decade or so has seen a number of serious attempts to remedy the situation, but these developments have been mostly concerned with either the drawing/extrusion or cross-rolling groups of processes. For historical as well as practical reasons, systematisation of work in these fields has been regarded as both more important and more productive because it offers a variety of possibilities. Eventually, these could be made applicable to more than one operation. This contrasts, of course, with the situation arising in, for instance, strip and plate rolling where the basic shape of the roll cannot be altered and improvements have to be sought through the medium of plant design. Forging constitutes a problem of its own since the almost infinite variety of shapes required precludes the possibility of formulating generalised solutions which are possible for axisymmetric processes.

5 Large-chip Metal Removal

D B Richardson

Contents

5.1 Large-chip processes 5/3

5.2 Cutting-tool geometry 5/6

5.3 Cutting-tool materials 5/10
- 5.3.1 High speed steel 5/10
- 5.3.2 Cast non-ferrous alloys 5/10
- 5.3.3 Sintered carbides 5/10
- 5.3.4 Ceramics 5/11
- 5.3.5 Cubic boron nitride 5/11
- 5.3.6 Diamonds 5/11
- 5.3.7 Limitations imposed by machine tools 5/11

5.4 Chip formation and cutting parameters 5/11
- 5.4.1 Influence of cutting parameters 5/13
- 5.4.2 Cutting fluids 5/13

5.5 Forces and power in metal cutting 5/14
- 5.5.1 Forces in turning processes 5/14
- 5.5.2 Forces and power in milling operations 5/15
- 5.5.3 Hot machining 5/17

5.6 Surface-finish considerations 5/17

5.7 Tool-life assessment 5/17
- 5.7.1 Crumbling of the cutting edge 5/18
- 5.7.2 Cratering 5/18
- 5.7.3 Plastic deformation of the tool 5/18
- 5.7.4 Thermal and mechanical shock 5/19
- 5.7.5 Flank wear 5/19
- 5.7.6 Catastrophic failure 5/19
- 5.7.7 Tool-life relationships 5/19

5.8 Economics of metal cutting 5/20
- 5.8.1 Minimum-cost production 5/20
- 5.8.2 Machining for maximum output 5/25
- 5.8.3 The influence of chatter 5/25

5.1 Large-chip processes

All the large-chip processes use cutting tools of defined geometry which are applied in a controlled manner to remove metal at a predetermined rate. The processes could be classified in many ways, but it is convenient to consider them in terms of the kinematics of the machine tools. With this in mind they have been separated into four main machine groups:

(1) turning (rotating work),
(2) shaping (reciprocating tool or work),
(3) milling (rotating tool), and
(4) drilling and boring (rotating tool).

Turning machines embrace the wide variety of lathes and vertical boring machines which can be controlled manually or automatically. Automatic control can be achieved using cams, sequential controllers, hydraulic copying devices or numerical programming. All machines in this group are capable of performing six basic operations, as shown in *Figure 5.1*. In addition, copying lathes and numerically controlled lathes can generate non-parallel forms by traversing the tool simultaneously in two planes.

Most turning processes use tools with a single cutting edge, where the cutting action is characterised by a relatively uniform section of material being presented to the cutting zone, resulting in a continuous chip when cutting ductile materials, or a repetitive form of short discontinuous chips when cutting brittle materials. Although the production of continuous chips indicates an efficient cutting action, the chip streamer itself presents disposal problems, frequently wrapping itself round the work piece, the cutter, or parts of the machine tool, creating a hazard to both the process and the operator. Chip breakers are extensively used to induce continuous chips to break into short lengths which are relatively safe and can be easily disposed of. These push against the underside of the chip and cause it to curl into a tight spiral, the free end of which strikes against the tool, and the resulting bending stress causes fracture. The earliest form of chip breaker, still extensively used with flat-top tools, consists of a hard wedge-shaped block of sintered carbide clamped to the rake face of the tool about 2 or 3 mm from the cutting edge (*Figure 5.2(a)*). The introduction of disposable sintered carbide inserts has allowed more complicated rake-face geometries to be used which act as built-in chip breakers (*Figure 5.2*(b)). Effective chip breaking is largely a matter of trial and error, being influenced by the feed, tool bluntness and cutting speed, as well as by the material being machined.

The development of new and improved cutting materials has enabled a 100-fold reduction in cutting time to be achieved since the beginning of the twentieth century. Unfortunately, the reduction in idle time—caused by the need for tool adjustment—and in the tool approach and retraction times before and after cutting has not been of a comparable order. In achieving lower production costs the emphasis has now rightly moved away from further reducing cutting time to attacking the disproportionately large amount of non-cutting time.

Disposable sintered cutting inserts are made to a high level of precision which allows them to be indexed or replaced in tool holders in a few seconds, usually without the need for sizing cuts. When all the cutting edges are worn the inserts are discarded, obviating the need for time-consuming regrinding which is common practice when using high speed steel tools. When resetting lathes between work batches, a substantial time saving can now be achieved by using preset tools mounted in holders which can be replaced as cartridges in the tool post. No doubt further improvements will be achieved in the development of cutting-tool materials, but any further reduction in manufacturing time achieved in the future will be determined mainly by reducing idle time, both by better tool-changing mechanisms and by improved machine tool design to facilitate chip disposal and to reduce the tool approach and retraction times.

The shaping group of machine tools produce chips by a relative linear motion between the cutting tool and the work. This group includes shaping machines, planing machines and slotting machines, all of which are used mainly for tool manufacture or maintenance work and have little application in modern production. They operate on a reciprocating principle, cutting on the forward stroke and idling on the return stroke. Although they have quick return mechanisms the cutting time is only of the order of half the reciprocating cycle time. Swarf disposal is usually no great problem due to the intermittent nature of the cut. *Figure 5.3* shows typical configurations of these three machine types, which have changed little in recent years.

Other machines in the shaping group are gear shapers and gear planers, outlines of which are shown in *Figure 5.4*. In gear

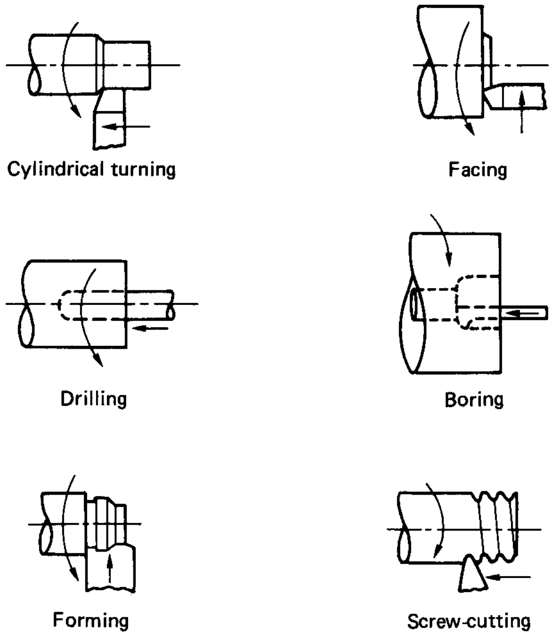


Figure 5.1 Basic lathe operations

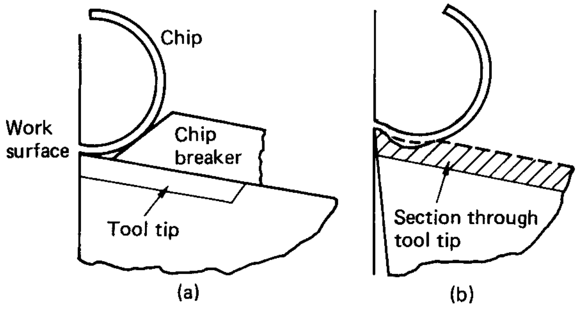


Figure 5.2 Chip breakers: (a) clamped; (b) built-in

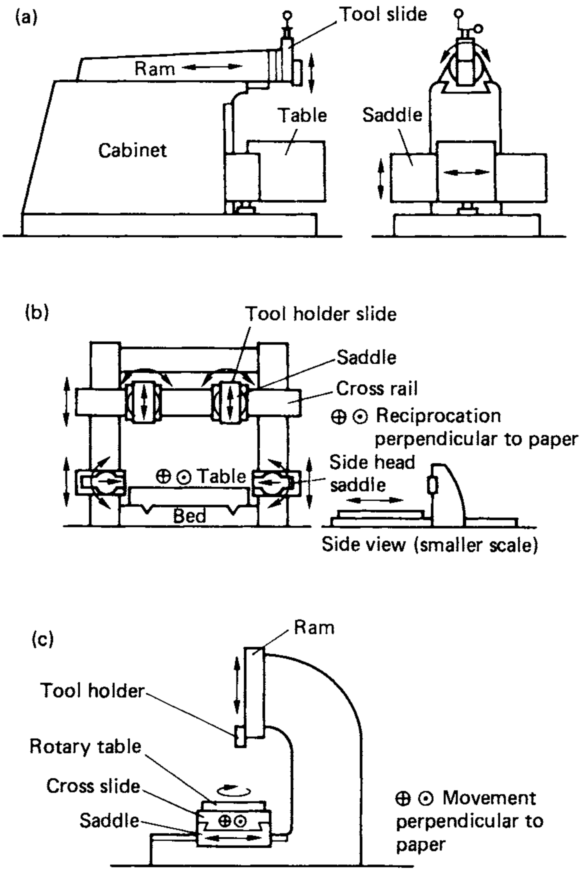


Figure 5.3 Outlines of reciprocating machines: (a) shaping machine; (b) planing machine; (c) slotting machine

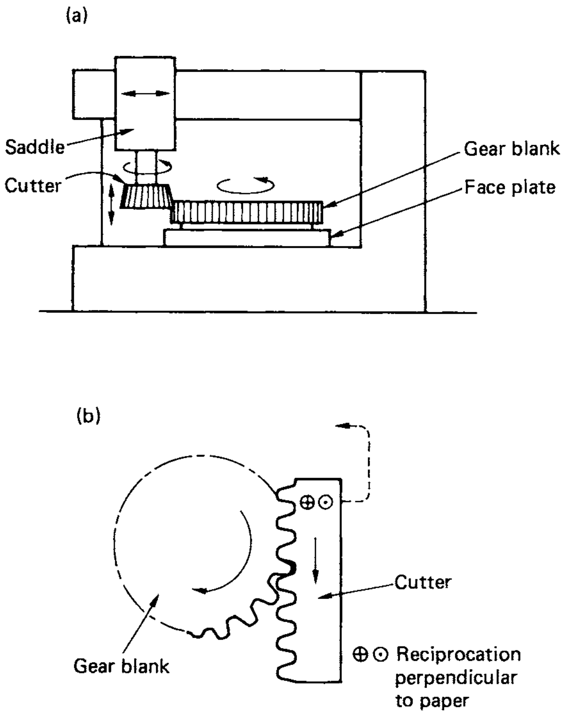


Figure 5.4 Kinematics of gear generating machines: (a) gear shaper; (b) gear planer

shaping the cutter resembles a side-relieved spur gear, the involute profile being generated by rapid reciprocation of the cutter whilst slowly revolving the cutter and gear blank in synchronism. Helical gears can be generated using a cutter with helical teeth and applying an appropriate helical motion to the spindles. Gear shaping is used for producing gears when hobbing would be impossible due, for instance, to a turned shoulder close to the involute profile. Gear planers have little modern use, particularly in a production environment. The cutter is in the form of a straight tooth rack, suitably relieved, and the gear is generated by reciprocating the cutter and moving the gear blank and the cutter at a constant speed. To enable a short rack to be used the cutter can be removed from the cut and indexed back at intervals.

Broaching machines also belong to the shaping group, but these produce the required form in a single pass. Internal broaching is for opening circular holes to produce non-circular forms. The cutter is a broach which has a number of cutting edges along its length and which is usually drawn, but sometimes pushed, through the hole by means of hydraulic pressure. Each cutting edge is larger than its predecessor by about 0.05–0.08 mm, so that the number of cutting teeth is determined by the form to be produced. Push broaching is limited to broaches with a small number of teeth which have a low length/cross-section ratio and which would not buckle under compression. Surface broaching is a more recently introduced technique, and is used as an alternative to milling for the production of external surfaces. Surface broaches are rigidly clamped to a machine slide and traversed against the component being machined, producing a surface in a shorter time than is required for milling, and usually giving a superior finish. Whereas internal broaching is usually the only feasible method for producing the desired shapes, surface broaching is an alternative to milling and is usually justified only if the quantities required are sufficient to offset the high equipment and tooling costs.

The milling group comprises a large range of manually operated or numerically controlled machines, many of which can perform operations such as drilling, reaming and boring as well as the accepted milling operations. Milling cutters generate surfaces either by means of cutting edges on the periphery or the face of the cutter. Peripheral milling is now seldom used for generating large plane surfaces, its main use being for machining slots or profiles. Although peripheral cutters can be fitted with carbide cutting edges the majority are of high speed steel and, except when used for machining the more exotic hard materials, will probably continue to be so in the foreseeable future. Frequently, such cutters also have shallow teeth on the cutter face, although these teeth usually contribute little to the total metal removed. A range of typical peripheral cutters is shown in *Figure 5.5*. Face milling cutters are essentially for generating plane surfaces. They are fastened to the end of stub arbors in the machine spindle and their configuration makes them suited to the use of specially designed carbide inserts (*Figure 5.6*). The cutting edges of both peripheral and face milling cutters are in contact with the uncut part of the component for, at most, half a revolution. Since the chip length is of the order of one-third of the length

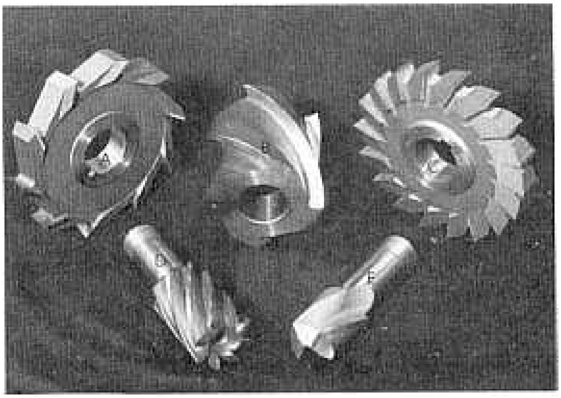


Figure 5.5 Peripheral cutters: A, high radial rake cutter; B, helical cutter; C, side and face cutter; D, end milling cutter; E, slot drill

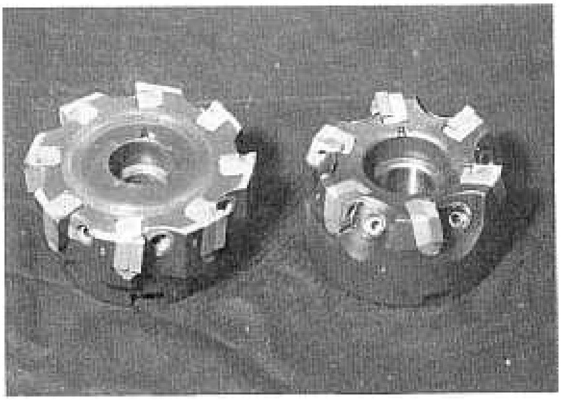


Figure 5.6 Face milling cutters: A, zero corner angle; B, 15° corner angle

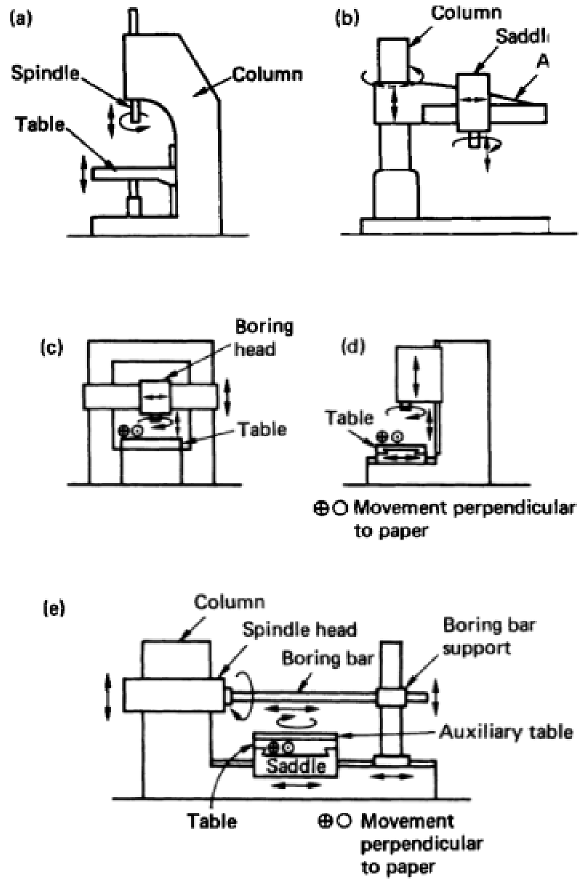


Figure 5.7 Machines in the drilling and boring group: (a) single spindle drill press; (b) radial drilling machine; (c) gantry type jig boring machine; (d) open front jig boring machine; (e) horizontal boring machine

of the uncut surface, the chips have a maximum length approximating to the cutter radius, so chip breaking poses few problems.

Gear hobbing machines also belong to the milling group. Hobs are in the form of a screw with a straight side rack-form thread, gashed to give cutting edges and relieved to provide cutting clearance. Gears are generated by synchronously rotating the blank and the hob, and the hob is fed parallel to the axis of the arbor on which the blank is mounted. When a number of identical gears is required, several blanks can be fitted to the same arbor and machined at a single pass.

The machines in the drilling and boring group can be subdivided into drill presses, radial drilling machines, jig boring machines and horizontal boring machines. They have rotating spindles which hold drills, stepped cutters, taps, reamers or single-point boring tools, and the cut is applied by feeding either the spindle or the work table. Typical configurations of cutters in this group are shown in *Figure 5.7*.

Increased metal-removal rates, made possible by the development of new cutting materials, have forced machine-tool manufacturers to design new machines capable of large ranges of spindle speeds and feeds. This has necessitated more powerful motors and structures having high rigidity to resist the increased cutting forces and to reduce the likelihood of self-induced vibrations, giving rise to chatter. With increased rates of metal removal, the problem of swarf disposal has become more acute. The magnitude of this problem can be realised by considering that swarf occupies about 100 times the volume of the metal removed, so a 10 kW motor running at full power can generate about 1 m^3 of mild steel swarf or about 3 m^3 of aluminium swarf per hour.

Modern high production metal cutting machines commonly cost in excess of £100 000. If they are amortised over a period of 5 years the depreciation cost of such a machine when used continuously on a double-shift basis is more than £5 per hour. It follows that such plant requires high utilisation, efficient programming to produce at optimal metal-removal rates, an effective system of tool management to reduce the non-cutting time and intelligent application of terotechnology to minimise lost time due to maintenance.

Where large-scale production justifies continuous or large-batch manufacture, the achievement of these objectives becomes a feasible possibility. Unfortunately, few products are marketed in such large quantities, and a large-batch approach usually results in uneconomically high stocks. In the 1970s, the British manufacturing industry became notorious for its in-

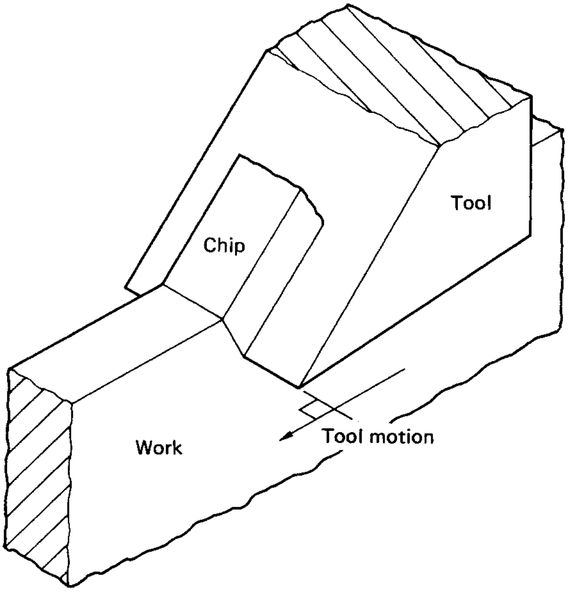


Figure 5.8 Orthogonal machining

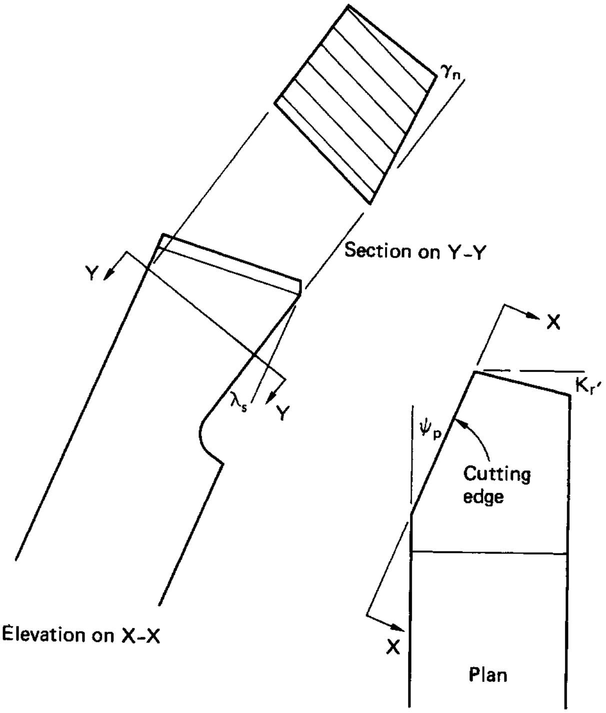


Figure 5.9 Angles in the normal rake system

flexibility and for its disproportionately high stocks, which resulted from the pursuit of large-batch policies. Present policies are directed towards small batches and this manufacturing philosophy has highlighted the need for rapid changeovers and for manufacturing systems accommodating large numbers of tools which can be called into use in response to the demands of small batches.

5.2 Cutting-tool geometry

Most of the research into chip formation has been based on orthogonal cutting, a simplified situation which is seldom met with in practice. The tool approaches the work with its cutting edge parallel to the uncut work surface and at right angles to the direction of cutting. To prevent end effects, the tool is wider than the work, as shown in *Figure 5.8*. Orthogonal cutting can be achieved only in a planing or shaping operation, although a close approximation can be obtained when turning on the end of a thin-walled tube.

In practice, the cutting tools usually approach the work obliquely and have rake angles in both directions on the rake face, together with a nose radius at the end of the cutting edge. The direction in which the chip flows across the tool surface is determined by this complicated geometry. BS 1296: 1972 defines the angles on single-point cutting tools in terms of the normal rake system (*Figure 5.9*), based on two coordinate rake angles.[1] The back rake or cutting edge inclination λ_s is measured parallel to the cutting edge in the vertical plane and the normal rake γ_n is measured in a plane at right angles to the cutting edge and perpendicular to the rake face. ψ_r is the tool approach angle and $\kappa_{r'}$ is the horizontal clearance angle, or the tool minor cutting edge angle. In addition, the tool is relieved to give vertical clearance angles of about 5°.

Other systems of tool nomenclature relate the rake angles to the coordinate axes of the tool shank, or to the cutting edge, measuring the angles in each case in the vertical plane. Although these systems are conceptually simpler, they are of little use in deducing the direction of chip flow. The British Standard relates to single-point tools but it can be applied also to multi-point tools and is generally preferable to the other systems.

Experimental work has shown that the direction in which the chip passes across the rake face of the tool can be expressed using Stabler's law.[2] This is best understood by reference to *Figure 5.10*, which shows an oblique shaping operation. The law simply states that $\gamma = \lambda_s$, where λ_s is the cutting-edge inclination and γ is the angle measured on the rake face between the normal to the cutting edge and the direction of chip flow. In this example the tool approach angle is zero, but the result is equally applicable for turning or face milling cutters where the tool approach angle is generally not zero.

The chip geometry and the principal cutting force are determined by a rake angle measured in the direction of chip flow. This is known as the 'effective rake' and is shown as β_e in *Figure 5.11*, which relates to a lathe tool, where

$$\begin{aligned}\sin\beta_e &= CE/AE = (BF + CD)/AE\\ &= BG/AE\cos\lambda_s + BD\sin\lambda_s/AE\\ &= BG\cos\gamma/AG\cos\lambda_s + (EG - FG)\sin\lambda_s/AE\\ &= \sin\gamma_n\cos\gamma/\cos\lambda_s + \sin\gamma\sin\lambda_s\\ &\quad - BG\tan\lambda_s\sin\lambda_s\cos\gamma/AG\\ &= \sin\gamma_n\cos\gamma/\cos\lambda_s + \sin\gamma\sin\lambda_s\\ &\quad - \sin\gamma_n\tan\lambda_s\sin\lambda_s\cos\gamma\end{aligned}$$

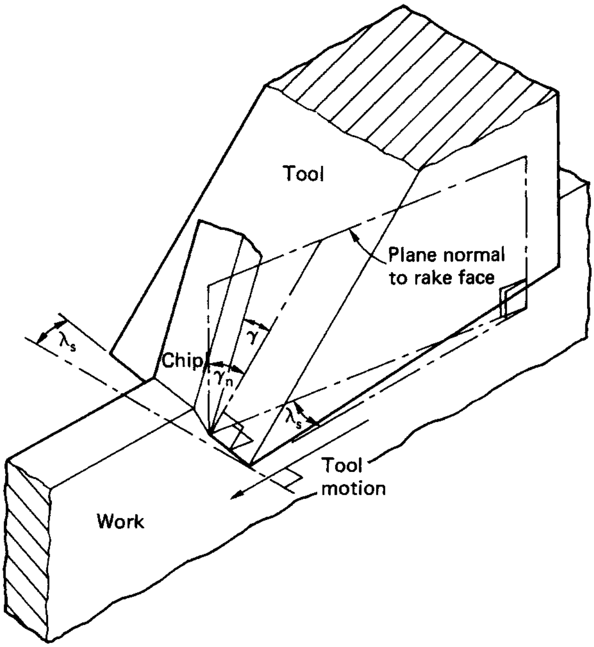


Figure 5.10 Oblique machining

Substituting $\gamma = \lambda_s$ (Stabler's law),

$$\sin \beta_e = \sin \gamma_n + \sin^2 \lambda_s - \sin \gamma_n \sin^2 \lambda_s$$
$$= \cos^2 \lambda_s \sin \gamma_n + \sin^2 \lambda_s \quad (5.1)$$

Figure 5.12 shows how β_e varies with changes in cutting-edge inclination and normal rake.

If the normal rake is not specified it can be calculated from the vertical rake angles coordinate to the cutting edge, as shown in *Figure 5.13*, from which it is seen that

$$\tan \gamma_n = BG/AB = BG/BF \cdot BF/AB$$
$$= \tan \beta' \cos \lambda_s \quad (5.2)$$

Figure 5.14 shows graphically how the normal rake is affected by variation in cutting-edge inclination and side rake.

The angles specified for a roller milling cutter are the radial rake and the helix angle (*Figure 5.15*), where the helix angle is the cutting edge inclination λ_s and the radial rake is β'.

Face milling cutters usually have a small nose radius at the end of the cutting edge, which helps to reduce the roughness of the machined surface. The cutting edge is usually inclined to the direction of travel by an angle θ, known as the corner angle, which is analogous to the tool-approach angle in *Figure 5.9*. This angle effectively increases the length of the cutting edge and reduces the uncut chip thickness t (*Figure 5.16*). By distributing the cut along a larger length, tool life can be increased. However, large corner angles can give rise to self-induced vibration, and they are normally restricted to values of $\leq 45°$.

The basic rake angles on face milling cutters are defined in the plane of rotation as the radial rake r_r and the axial rake r_a. To obtain the cutting edge inclination and the normal rake the following transformations (based on *Figure 5.17*) are necessary:

$$\tan \lambda_s = (CC' - BB')/B'C'$$
$$= (AC' \tan r_a - AB' \tan r_r)/B'C'$$
$$= \cos \theta \tan r_a - \sin \theta \tan r_r \quad (5.3)$$

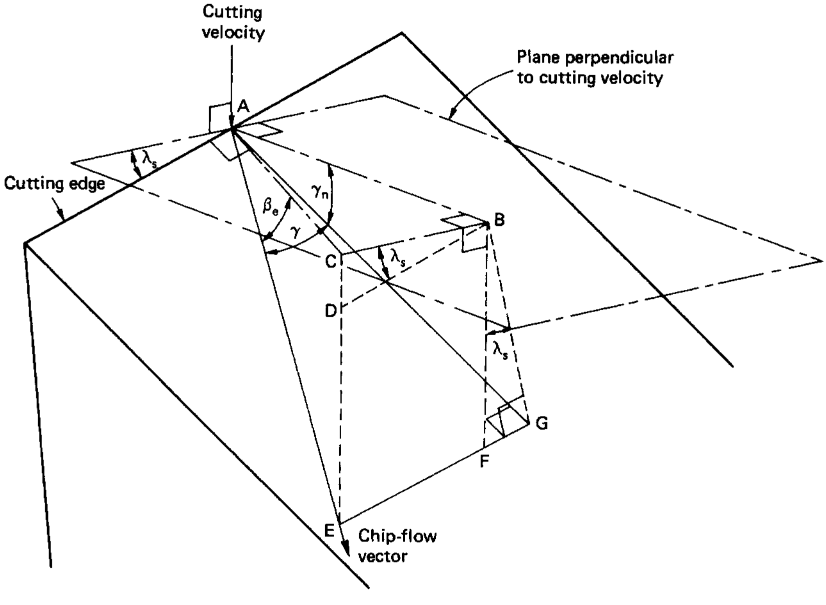


Figure 5.11 Angles in oblique machining

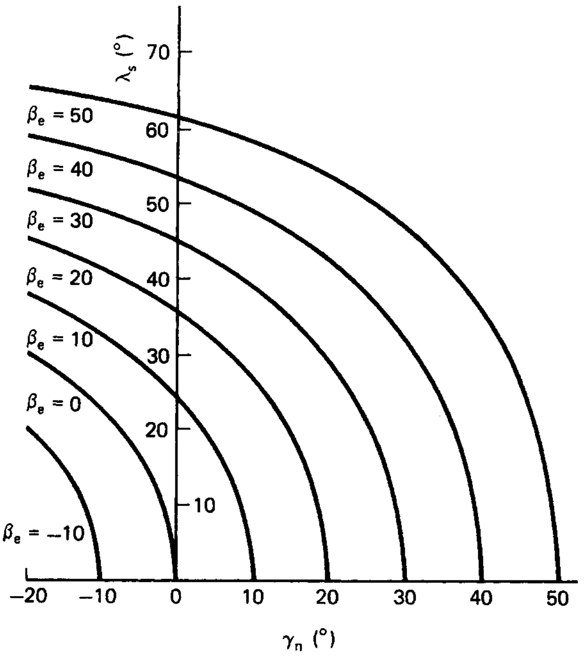


Figure 5.12 Effective rake relationships

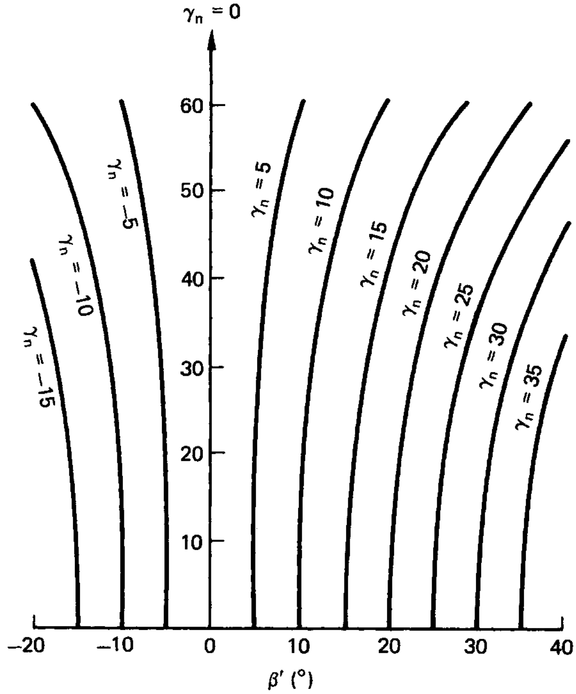


Figure 5.14 Rake-angle relationships for oblique cutting

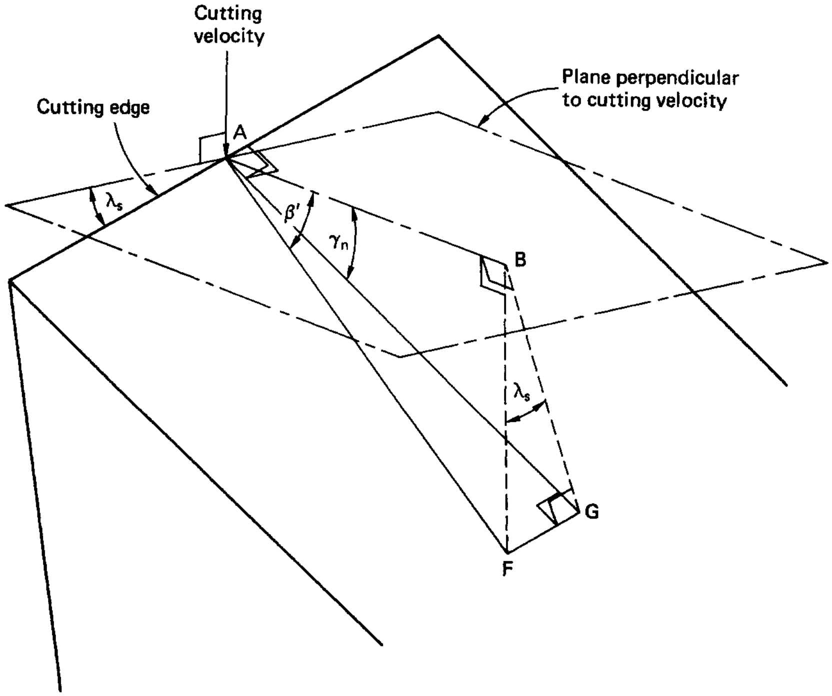


Figure 5.13 Back rake and normal rake relationship

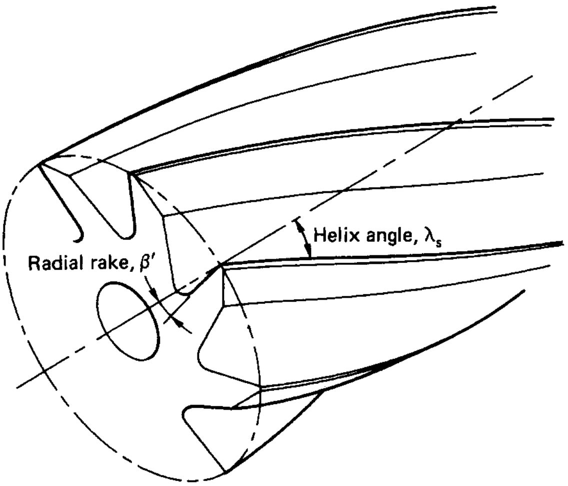


Figure 5.15 Rake angles on a helical milling cutter

The side rake β' is evaluated by reference to *Figure 5.17(b)* as follows:

$$\tan \beta' = EE'/DE'$$
$$= (DF' \tan r_a + EF' \tan r_r)/DE'$$
$$= \sin \theta \tan r_a + \cos \theta \tan r_r \qquad (5.4)$$

The normal rake γ_n is evaluated from equation (5.2):

$$\tan \gamma_n = \tan \beta' \cos \lambda_s$$

and the effective rake angle is evaluated from equation (5.1):

$$\sin \beta_e = \cos^2 \lambda_s \sin \gamma_n + \sin^2 \lambda_s$$

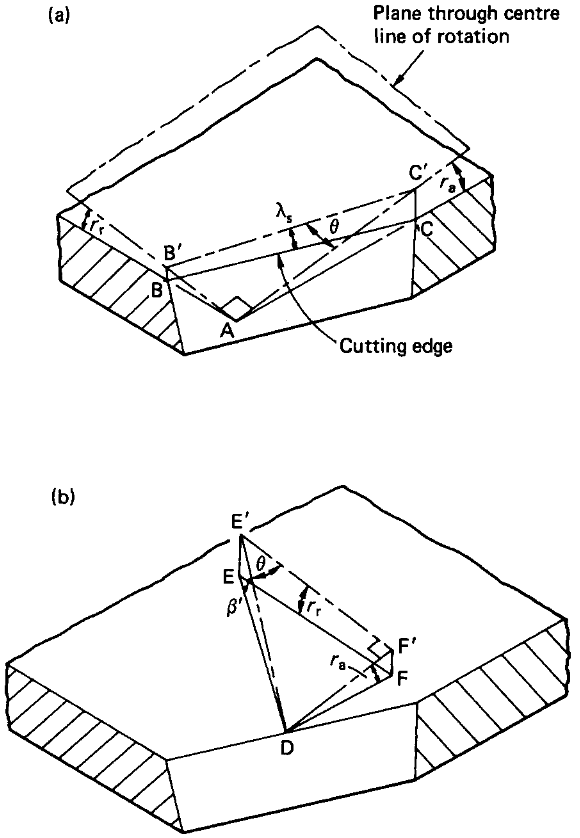


Figure 5.17 Geometry of a face milling cutter

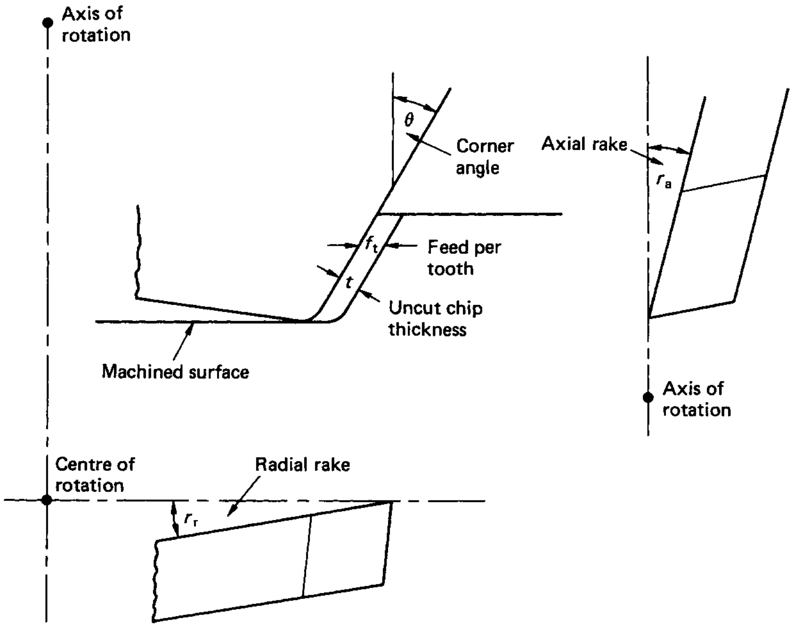


Figure 5.16 Basic angles on a face milling cutter

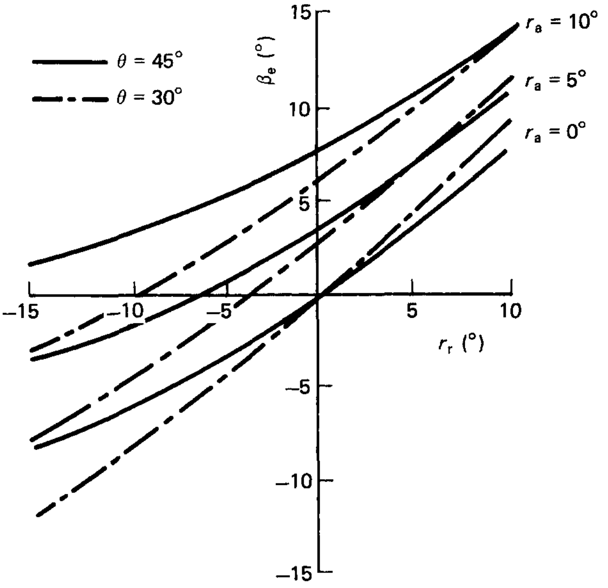


Figure 5.18 Cutting-angle relationships on a face milling cutter

Figure 5.18 shows families of curves for cutters having corner angles of 30° and 45°, where β_e is plotted against r_r for three values of r_a. It can be seen that the benefits of negative rake angles can be obtained by selecting suitable combinations of angles whilst retaining the lower axial thrust associated with positive rakes. In addition, as demonstrated in Section 5.7, a large corner angle gives a longer tool life.

5.3 Cutting-tool materials

The range of available cutting-tool materials has increased rapidly in recent years due to the development of more difficult to machine materials and to the demand for higher productivity. Any attempt to summarise these materials is unlikely to be completely successful due to the rate at which improvements and innovations are occurring. However, this is no excuse for ignoring the published state of the art at the time of writing.

When selecting a tool material for a particular application, it is necessary to measure its rating against the following list of properties, some of which are mutually opposed. For instance, in most cases hardness and impact resistance of competing tools tend to be inversely related. The essential properties are:

(1) high hardness at elevated temperatures,
(2) high compressive strength,
(3) adequate impact resistance,
(4) insusceptibility to large variation in local temperature,
(5) chemical inertness at working temperatures,
(6) low interface friction, and
(7) good abrasion resistance.

Temperatures at the chip–tool interface can be about 1000°C when machining steel, and can be considerably higher when machining some of the more exotic materials, particularly if heat is applied to help soften the work material. As the cutting temperature is largely dependent on cutting speed, it follows that the pursuit of higher productivity creates increasing demands on the high-temperature properties of the tool materials.

5.3.1 High speed steel

High speed steels are likely to continue to be used in the foreseeable future for many applications such as drilling, reaming, tapping and dieing, forming, broaching and milling due to the ease with which they can be shaped in both the soft and hardened state. Typically, they consist of carbon steel alloyed with tungsten or molybdenum, together with additions of chromium, vanadium and cobalt. The alloying elements raise the temperature at which tempering occurs, allowing high speed steels to be used at temperature up to about 650°C. Their hardness is limited to 750 H_V (Vicker's hardness number), which is adequate for machining most of the common metals, including alloy steels in their unhardened forms. Cutting speeds are necessarily limited to prevent excessive rise in temperature. When machining mild steel, cutting speeds of about 1.5 m s^{-1} are possible if a plentiful supply of coolant is provided. A recent development is the coating of high speed steel drills with a deposit of about 3–5 μm of titanium nitride which allows rotational speeds to be increased, resulting in a 50% increase in penetration rate and longer tool life. With a few exceptions of this sort, it is unlikely that high speed steels will ever again pose a threat to the supremacy of sintered carbides for heavy-metal removal.

Despite their relatively low hardness and susceptibility to softening at high temperatures, high speed steel tools are tougher than most of the competing materials, enabling them to be used for interrupted cuts without fear of fracture. They can also be reground, giving a number of cutting lives before they must be finally discarded.

5.3.2 Cast non-ferrous alloys

These alloys consist of cobalt, chromium, tungsten and carbon. Although less versatile than high speed steels, they enable cutting to be performed at higher temperatures. Their main use is for drilling, where their superior hardness at elevated temperatures is an advantage when the application of fluids is frequently a problem. Their cutting performance is generally superior to high speed steel, but inferior to that of sintered carbides, so their use is unlikely to grow in popularity.

5.3.3 Sintered carbides

The introduction of sintered carbides for cutting has been the most important single contribution to increased productivity during the past 50 years. Sintered carbides are essentially cermets, which consist of hard carbide ceramic particles embedded in a metal matrix. Early carbide tools were usually of tungsten carbide and cobalt. Their brittleness encouraged the use of negative rakes to promote compressive stresses and restricted their use to continuous cutting. Subsequent improvements in the sintering process, and the introduction of alternative ceramic and metallic components, has enabled a range of carbide tools to be produced which withstand the rigours of thermal and mechanical shock, making them suitable for interrupted cutting in turning and milling.

The first generation of carbide tools consisted of sintered tips brazed to steel shanks. This arrangement provided a fairly rigid cutting system and the tools could be reground when blunt. More recently, the brazed tip has been almost completely superseded by disposable tips, which are mechanically clamped into steel tool holders. These tips are polygonal, with three or more cutting edges which can be indexed when worn

to expose new edges. The negative-rake types can be inverted to double the number of edges. They are a throw-away concept, regrinding being uneconomical. Their introduction has forced a reappraisal of metal cutting economics, making tool lives of the order of 15 min a desirable objective. Practical cutting speeds are about three times as great as could be achieved with high speed steel, 5 m s^{-1} being a typical maximum when machining mild steel.

Depending on their composition, the hardness of sintered carbides varies between 1200 and 1600 H_V, impact resistance being obtained at the cost of hardness. Thermal softening becomes a problem only if the cutting temperature is allowed to exceed 1000°C, enabling most cutting operations to be performed without the use of a coolant. Improvements in cutting performance are now obtained by applying a variety of vacuum deposited coatings to the surface of the carbide tools. Although the coatings are usually only 1.0–5.0 μm thick, when used singly or in combination they can enhance tool life by a factor of about 2. Thus, a coating of titanium carbide reduces wear due to cratering on the rake face, aluminium oxide increases the surface hardness, and titanium nitride reduces rake face friction.

5.3.4 Ceramics

Sintered ceramic tools based on aluminium oxide (Al_2O_3) have been available for more than 30 years. Their brittleness and poor thermal shock resistance can be improved by additions of zirconium oxide and titanium carbide, but they are generally unsuited to interrupted cutting. Having hardness values in the range 1500–1800 H_V, which does not significantly decrease over the normal metal cutting temperature range, they are suitable for machining hardened steels and chilled cast irons. They are similarly suitable for heat-assisted machining of hard materials such as the nimonic alloys and Stellite. However, their brittleness has proved a severe limitation for general-purpose machining of steels, an area in which their high metal removal potential would have been an advantage.

Mixed ceramics, based on carboxides with dispersed titanium carbide, have achieved better impact resistance without significant loss of hardness. They can be used at high cutting speeds of the order of 15–20 m s^{-1} when operated at low feeds, making them suitable for finish turning and finish milling hard materials.

A recent addition to the range of commercially available ceramic cutting tools uses silicon nitride (Si_3N_4) with various levels of aluminium and oxygen substitution. The silicon nitride ceramics have good resistance to thermal and mechanical shock, enabling them to be used for discontinuous cutting. They can be operated at higher cutting speeds than carbides together with higher feeds than can other types of ceramics. Their main applications to date are for rough turning and rough milling grey cast iron and for turning nickel alloy steels. Due to their high thermal shock resistance they can be used to cut dry or with coolant, the latter method being inadvisable with other ceramics. Chemical reaction occurs at cutting temperatures, causing rapid wear when machining most steels. Thus, at present, there is little likelihood of silicon nitride supplanting carbides in this important area of manufacture.

5.3.5 Cubic boron nitride

Polycrystalline cubic boron nitride is another comparatively recently introduced addition to the range of metal-cutting materials. It has hardness considerably in excess of that of ceramic tools and retains its hardness at temperatures well in excess of 1000°C. The main application is for machining hard ferrous materials at very high cutting speeds, giving a surface finish comparable to that obtained by grinding. It has relatively good impact resistance, allowing it to be used for interrupted cutting. It is also of use for hot machining of refractory metals such as Stellite, where a 90% reduction in machining time compared with carbide has been claimed. Due to the high cost of cubic boron nitride tools, manufacturers do not claim great cost savings, but the time saving is very significant.

5.3.6 Diamonds

Diamond is the hardest material known to man and it has found a limited cutting application where this is an important attribute. Natural diamonds, brazed to steel holders, have been used for many years for producing fine finishes on copper and aluminium. Being monocrystalline, natural diamonds have planes of weakness which render them unsuitable for anything but fine finishing cuts.

The production of polycrystalline synthetic diamonds has extended the usefulness of diamonds by improving their impact resistance. Polycrystalline synthetic diamond cutting tools are now used extensively in machining abrasive aluminium–silicon alloys, fused silica, and reinforced plastics. They are chemically reactive at high temperatures, so they are of little use for machining ferrous materials.

5.3.7 Limitations imposed by machine tools

Industry has, until recently, been very reluctant to replace machine tools whilst they continue to perform the function for which they were purchased. Machine-tool manufacturers market machines which adequately utilise the cutting tools available at the time of purchase. Inevitably, with the rapid development of new cutting materials, the existing machine tools cease to provide a service which uses the cutting tools in an economical manner.

Self-induced vibration, giving rise to chatter, is undesirable in any cutting operation. It is particularly undesirable when using brittle cutting tools where catastrophic failure can become a very real possibility. The problem of self-induced vibration becomes more acute as metal removal rates increase at high cutting speeds. Ideally, resonant frequencies should be as high as possible, but this requires high structural stiffness and low mass. Unfortunately, dynamic stiffness tends to be directly related to mass, so a simple scale-related solution does little to reduce chatter. The solution, if it exists, lies in structural redesign to enhance stiffness without a proportional increase in mass.

Optimal cutting speeds using modern cutting tools require a large increase in the rotational speed of spindles and a corresponding increase in input power. Manufacturers of machine tools recognise this need, which is reflected in their latest designs, together with improved provision for handling the greater volumes of swarf which are produced.

5.4 Chip formation and cutting parameters

It is convenient to assume that a chip shears across a plane and that this is the sole mechanism of deformation in the cutting process (see *Figure 5.19*). Much useful research in metal cutting has been developed using this simple analogy. Unfortunately, it is a gross oversimplification due to two factors: the effects of temperature and strain rate on the yield behaviour of work materials. The shear plane analogy, literally interpreted, implies an infinite rate of shear strain at the shear plane, giving

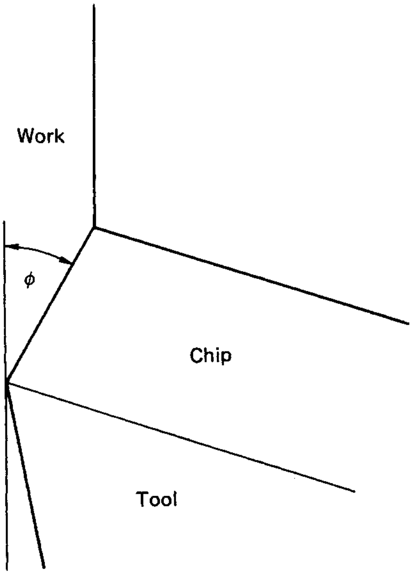


Figure 5.19 Ideal chip formation

rise to an instantaneous change in velocity and, presumably, an infinite force to effect it.

In practice, primary shearing occurs in a roughly triangular zone originating from the cutting edge. Friction effects along the initial length of the rake face cause a secondary deformation zone due to retardation of the underside of the chip (*Figure 5.20*). When the clearance flank of the tool wears, a further deformation zone may occur on the cut surface of the work. This zone is generally small and can be ignored, unless the surface integrity of the component is of particular importance. Measurement of the deformation of fine square grids scribed or photographed on the sides of the work piece has shown that, although the deformation zone departs dramatically from the shear plane concept, the very high strain rates are concentrated in a narrow band which follows closely the hypothetical shear plane.[3] In the following analysis of chip formation, the shear plane analogy is used extensively to demonstrate how the parameters of cutting speed and feed affect the chip root configuration.

A further complication affecting the understanding of chip formation is the existence of a built-up nose, usually referred to as a 'built-up edge', on the underside of the chip adjacent to the cutting edge. This results from tool bluntness and rake face friction, and its formation can be visualised as follows. Assume that the tool has a slight wear radius along the cutting edge. A particle of work material at the bottom extremity of the shear plane reaches this radius which acts as a stagnation point, the particle being constrained equally to join the cut work surface and the underside of the chip. Succeeding particles, impinging on the first, become stationary and a built-up nose is initiated. Adjacent particles slightly higher on the shear plane do not stagnate, but proceed along the rake face where the high compressive stress causes a friction force of sufficient intensity to initiate sticking friction. In this way a velocity gradient is created, varying from zero at the rake face to the full chip velocity at a line located part way into the chip. Eventually, an equilibrium is achieved where the effect of the temperature gradient produces a shear flow stress large enough to prevent further secondary deformation. This determines the boundary of a persistent, if unstable, built-up edge (*Figure 5.21*). The front of the built-up nose is constantly breaking away and being replaced, the debris being transported by the cut surface, producing a characteristically poor finish. Elevated temperatures at the interface between the tool and the built-up edge can cause diffusion welding to occur, leading to tensile failure of the surface on carbide tools.

Obviously-built-up edge is an undesirable phenomenon which should be avoided if possible. It occurs commonly when using high speed steel or carbide tools, but causes little problem with ceramics. Although the reason for this is not fully known, it is probable that the very low thermal conductivity of ceramics causes a high interface temperature and a

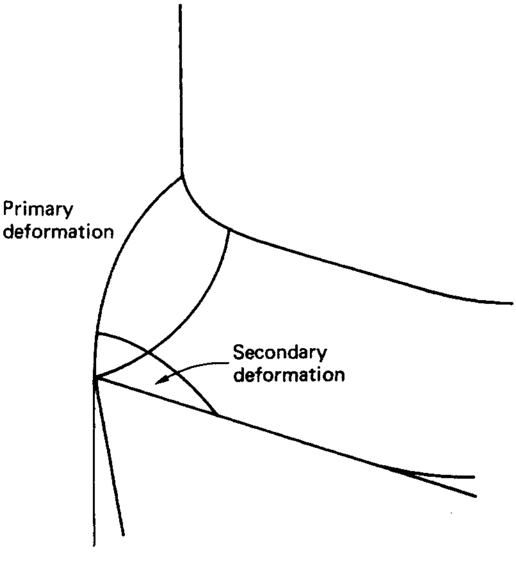


Figure 5.20 Actual chip formation

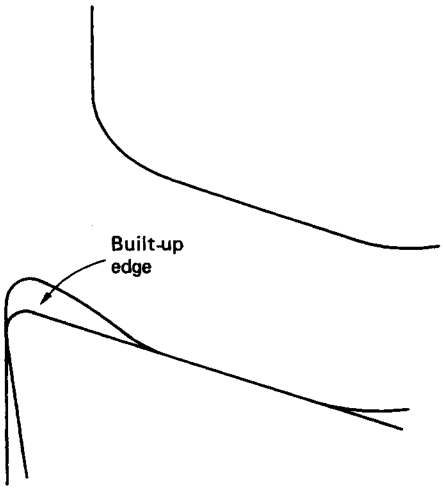


Figure 5.21 Typical built-up edge formation

consequent reduction in shear flow stress on the underside of the chip which, in turn, localises the secondary deformation. When cutting steel at impractical cutting speeds below about 0.050 m s^{-1}, there is no significant built-up edge. Such an edge assumes troublesome proportions at cutting speeds between about 0.2 and 1.2 m s^{-1} and then ceases to be significant at cutting speeds above about 2.0 m s^{-1}. The effect can be observed by performing cutting tests at increasing speed, where surface roughness improves progressively prior to a dramatic improvement at speeds in excess of about 2.5 m s^{-1}. Cutting speeds of this order are economically possible when cutting mild steel with carbides, but may not be economical when cutting some high carbon and alloy steels.

Some aluminium alloys promote a built-up edge which persists at very high cutting speeds. This accounts for the fact that, although these alloys are easily machined, it is frequently extremely difficult to produce a good surface on them.

5.4.1 Influence of cutting parameters

Heat generated in the primary shear zone is mainly carried away by the chip. However, some can be back-conducted into the work piece. This raises the temperature ahead of the shear plane and reduces the shear flow stress of the incoming material. In turn, the shearing action occurs at an earlier point, causing the shear plane to rotate and reducing the shear angle ϕ in *Figure 5.19*.

Since heat conduction is time dependent, the amount of heat conducted to the body of the work material at low cutting speeds is likely to be greater than at high speeds, even allowing for the greater total amount of heat generated at higher speeds. This belief is consistent with the observed fact that a surface machined at high speeds is usually cooler to the touch than one machined at low speeds. Furthermore, chips produced at high speeds are noticeably thinner than those produced at low speeds, a result of shear plane rotation. The principal cutting force decreases as cutting speed increases, due to a shorter shear plane, although this effect is frequently obscured by the effect due to a built-up edge.

When cutting with low feeds the shear angle is small, but it increases as the feed increases. This, no doubt, is also attributable to the effect of heat conduction, although the argument is more tenuous than for speed variation.

Intuitively, the maximum shear angle could be half the angle between the work surface and the rake face. In practice, it is always appreciably smaller. Sintered tools are operated over a small range of rake angles, varying between about +10° and −10°. When using negative rakes the shear angle is correspondingly reduced, giving very low angles when cutting at low speeds. However, negative rake cutting is essentially a high-speed operation, and at high cutting speeds the shear angle is not much less than that using positive rakes.

If the shear-angle concept is modified to assume that primary deformation occurs across a narrow band (*Figure 5.22*), the primary shear strain γ depends on the rake angle β_e and the shear angle ϕ. A particle entering the shear zone at O changes direction under the influence of shear and leaves the shear zone at B:

$$\gamma = AB/OC = (AC + BC)/OC$$
$$= \cot \phi + \tan (\phi - \beta_e) \quad (5.5)$$

It can be seen that primary shear strain, and consequently the energy of primary deformation, are reduced if the effective rake angle β_e and the shear angle ϕ are large.

The ratio of the uncut chip thickness t divided by the actual chip thickness t' is generally denoted by r_t (*Figure 5.23*),

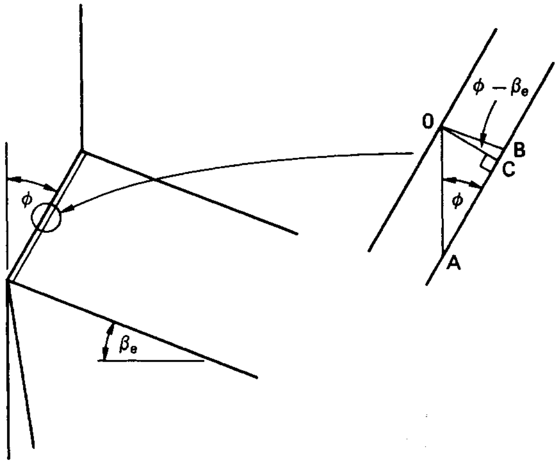


Figure 5.22 Rectangular shear zone model

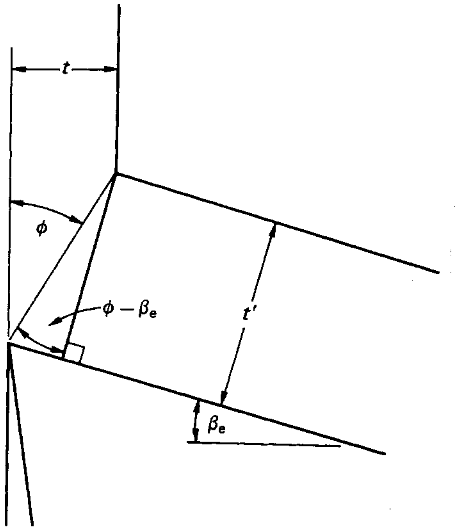


Figure 5.23 Chip thickness relationship

where it is easily deduced that:

$$\tan \phi = (1 - r_t \sin \beta_e)/(r_t \cos \beta_e) \quad (5.6)$$

High values of r_t, i.e. thin chips, are sometimes taken as an indication of efficient cutting, giving rise to reduced primary shear strain and, consequently, to lower cutting forces.

5.4.2 Cutting fluids

Cutting fluids are used for three main purposes: as a lubricant at low cutting speeds, to cool the tool and work, and to assist in clearing the swarf.

At cutting speeds in excess of about 0.7 m s^{-1}, there is little noticeable lubricating effect. Below this speed extreme pressure (EP) mineral oils containing sulphur or chlorine additives can be used to reduce friction in the latter stages of chip–tool

contact due to the formation of low-shear-strength sulphides or chlorides. The balance of the cutting forces is affected, giving rise to a larger shear angle and reduced contact length, and encouraging the thinner chip to curl, making for more effective chip breaking. High compressive stresses near the tool point prevent lubricant penetrating in this area, so the lubricating effect is limited to the latter part of the chip–tool interface where sliding friction occurs. The chemical reaction giving rise to low-strength compounds is both temperature and time dependent. Hence, at higher cutting speeds the lubricant rapidly loses its effectiveness.

High speed steel cutting tools start to soften at temperatures above about 650°C. When using these materials, the cooling effect of cutting fluids enables the tools to be operated at higher cutting speeds than would be possible when cutting dry. Water soluble oils, having high specific heat and good metal wetting properties, are better coolants than the mineral oils when used as lubricants. With carbides and ceramics the poor resistance to thermal shock makes the use of cutting fluids inadvisable, except in special circumstances. Fortunately, these materials can be used satisfactorily at high temperatures, and coolants are therefore not usually required.

The purely mechanical function of using cutting fluid to assist swarf disposal is sometimes of prime importance. An example of this application is in deep-hole drilling where cutting fluid is pumped to the cutting edges at high pressures of about 6 N mm^{-2}.

5.5 Forces and power in metal cutting

Most lathes and milling machines lack the power to exploit the cutting tools in an economical manner. This shortcoming is usually aggravated by a natural trepidation on the part of operators to run machine tools near their power limits, for fear of stalling the drive motor. Surprisingly, few machine tools are fitted with wattmeters, so the operator usually has no idea how near he/she is to causing an overload.

Among the more sophisticated numerically controlled machines, very few are fitted with adaptive control devices which cause feed or cutting speed to respond to excessive power demands. The vast majority operate from a predetermined programme which has been based on safety considerations, where the power requirements are well within the rated output of the motors.

There is an ill-founded belief that the cutting forces, and hence the power required, increase significantly as the tools wear, increases of 40% sometimes being quoted. In fact, cutting power seldom increases by more than 10% over the life of the tools.

At rated power, transmission losses usually account for about 30% of the input power, with a correspondingly greater percentage loss when operating at lower energy levels. Transmission losses are higher when the machine tool is cold, and drop significantly over the first half hour of operation. It is desirable, therefore, to record the transmission power over the full range of cutting speeds and feeds on a machine tool in both the cold and warmed-up conditions. Only then is it possible to know the available power which can be used for cutting.

5.5.1 Forces in turning processes

Although a knowledge of cutting forces is desirable to prevent excessive structural loads, the main reason for requiring such knowledge is to use it as a basis for estimating power. The power in watts is simply the product of the peripheral speed in metres per second and the tangential cutting force measured in newtons.

Due to the formation of built-up edge when cutting steel at low speeds, forces on the tool vary in an unpredictable manner, but above about 2.5 m s^{-1} become relatively constant when built-up edge ceases to have a significant effect. At cutting speeds below about 0.7 m s^{-1}, the lubricating effect of cutting fluids can do much to inhibit build up, and on the rare occasions when such low speeds are used the cutting force can often be reduced by this means. When using sintered tools it is usually possible to operate at cutting speeds high enough for the forces to be considered constant.

As would be expected, the effective rake has a significant effect on cutting force but, as mentioned in the previous section, the shear angle disparity between positive and negative rake tools becomes less at high cutting speeds. *Figure 5.24* shows approximately how the principal cutting force varies with effective rake at both high and low cutting speeds. High speed steel tools typically operate at the lower speed, where the cutting force increases by about 90% as the rake decreases from +30° to zero. Carbide tools, which operate typically at the higher speed, experience a rise in cutting force of only about 15% across the range of rakes from +10° to −10°.

This would seem to indicate a decided advantage in using high rakes when cutting at low speeds. However, tool-life considerations militate against high rakes when machining hard materials, because the restricted heat conduction path from the cutting edge leads to high interface temperatures. Typical rake angles for high speed steel tools are shown in *Table 5.1*.

As discussed previously, the shear angle increases as the uncut chip thickness increases, so that the shear plane length does not increase linearly with feed. This causes the cutting force to be related non-linearly to feed f, or to uncut chip

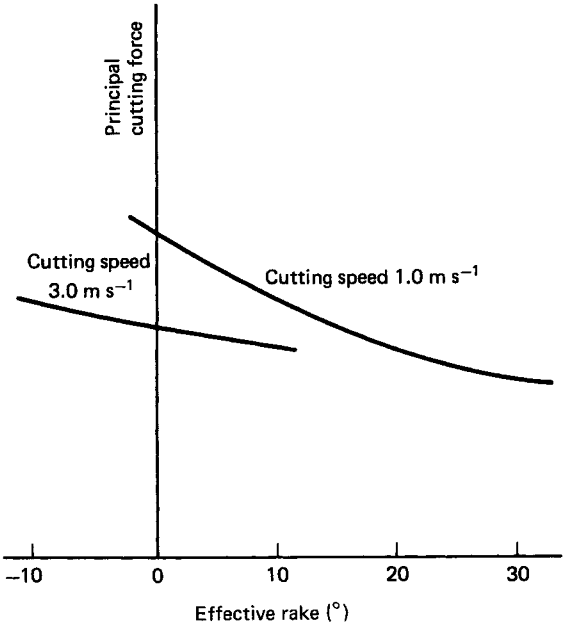


Figure 5.24 Variation of cutting force with rake and cutting speed

Table 5.1 Typical rake angles for high speed steel tools

Work material	*Rake angle* (°)
Aluminium and soft alloys	30–40
Free cutting steels	20–25
Medium carbon steels	10–15
Hard alloy steels, hard cast iron	0–5

thickness t. When cutting orthogonally, where $f = t$,

$$F = Ct^x w \tag{5.7}$$

where F is the principal cutting force (in newtons), C is a material constant at a given cutting speed and rake angle, t is the uncut chip thickness (in mm), w is the width of the cut (in mm) and x is a constant for the material being machined.

The exponent x varies slightly with rake angle, but for practical purposes can be assumed to be 0.85 when machining the whole range of steels and about 0.7 when machining cast irons. Average values of C when using positive or negative rake carbide tools at typical cutting speeds are given in *Table 5.2*.

In the more practical case of oblique cutting, equation (5.7) must be modified to allow for the tool approach angle and cutting edge inclination. The plan geometry of chip flow for a pointed tool, with an approach angle ψ_r and cutting edge inclination λ_s, is shown in *Figure 5.25*, where the angle δ is angle BAC in *Figure 5.11*.

The equation for cutting force becomes modified as

$$F = Ct'^x w' \tag{5.8}$$

where t' is the uncut chip thickness in the direction of chip flow, and w' is the length of cutting edge in engagement.

Table 5.2 Average values of C at typical cutting speeds for positive and negative rake carbide tools

Work material		*Specific angle* (°) +10	−10	*Specific power*
Free cutting carbon steel	120HB	980	1080	2.6
	180HB	1190	1310	3.1
Carbon steels	125HB	1620	1780	4.2
	225HB	2240	2470	5.8
Nickel–chromium steels	125HB	1460	1610	3.8
	270HB	2100	2310	5.4
Nickel–molybdenum and chromium–molybdenum steels	150HB	1600	1760	4.1
	280HB	1960	2160	5.1
Chromium–vanadium steels	170HB	1820	2000	4.7
	190HB	2380	2620	6.2
Flake graphite cast iron	100HB	635	700	1.6
	263HB	1330	1470	3.5
Nodular cast iron	Annealed	1110	1220	2.9
	As cast	1240	1370	3.2

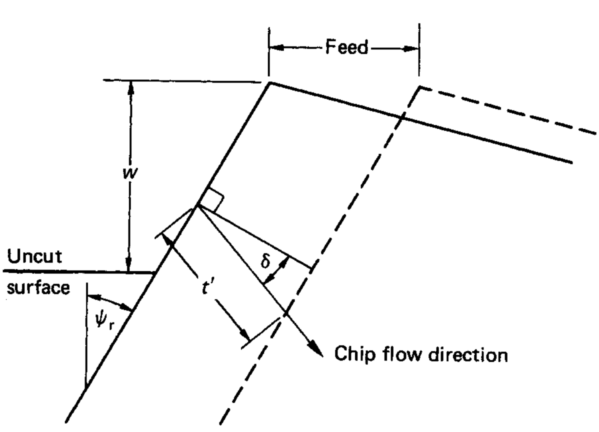


Figure 5.25 Plan geometry of a pointed tool

It can be shown that,

$$\cos \delta = \cos \lambda_s \cos \gamma_n / \cos \beta_e$$

and

$$t' = f \cos \psi_r / \cos \delta$$

$$= f \cos \psi_r \cos \beta_e / (\cos \lambda_s \cos \gamma_n) \tag{5.9}$$

where f is the feed per revolution, and ψ_r is the tool approach angle.

Allowing for the effect of tool approach angle and cutting edge inclination:

$$w' = w/(\cos \psi_r \cos \lambda_s) \tag{5.10}$$

Expressed in terms of feed and width of cut, equation (5.8) becomes

$$F = Cf^x w(\sec^{1-x} \psi_r \cos^x \beta_e \sec^{1+x} \lambda_s \sec^x \gamma_n) \tag{5.11}$$

Considering a practical range of values for the angles used in industrial practice, the force calculated from equation (5.11) seldom differs from that calculated from equation (5.7) by more than 6%. Allowing for the effect of tool wear, it would seem reasonable to multiply the result obtained from equation (5.7) by a factor of 1.2 to give a simple and slightly overvalued estimate of cutting force.

However, when applied to the peripheral milling operation, where helix angles exceeding 45° are not uncommon, the effect of obliquity significantly affects the cutting force and the cutter geometry cannot be ignored.

5.5.2 Forces and power in milling operations

Peripheral milling removes metal by means of teeth on the circumference of the cutter. It is seldom used to produce large flat surfaces, which are more effectively generated by face milling. Mostly, peripheral milling cutters are used for end milling slots or for producing slots or stepped surfaces by using one or more horizontally mounted side and face cutters or helical slab milling cutters.

End milling seldom requires the rated power of the drive motor. The limiting factor is usually the maximum recommended feed per cutting tooth which will prevent damage to the cutter. Horizontal peripheral milling, however, can be limited by the power of the drive motor, and it is useful to consider the way in which the cutting parameters affect the power required.

Although the changing chip section precludes the possibility of orthogonal milling, it is possible to consider the instantaneous situation when using a zero helix cutter as quasi-orthogonal. Due to the motion of the work the profile of the surface generated by a cutter tooth is not strictly circular, but there is not significant loss of accuracy in assuming it to be so. It then follows, from *Figure 5.26*, that,

$$\cos \theta_m = 1 - d/R \tag{5.12}$$

where d is the depth of cut (in mm), and R is the radius of the cutter (in mm).

Figure 5.27 shows the uncut chip geometry and considers the instantaneous uncut chip thickness when the cutting edge has rotated through an angle θ from the vertical,

$$t \simeq f_t \sin \theta$$

where f_t is the feed per cutting tooth (in mm) and t is the instantaneous uncut chip thickness (in mm).

From equation (5.7),

$$F = Ct^x w$$

and the work done in removing one chip is

$$W = RwC \int_0^{\theta_m} t^x d\theta/1000 \text{ watts}$$

$$\simeq RwCf_t^x \int_0^{\theta_m} (\sin \theta)^x d\theta/1000 \text{ watts}$$

Since θ is usually less than 1 rad,

$$W \simeq RwCf_t^x \, \theta_m^{1+x}/[1000(1 + x)] \text{ watts}$$

Substituting from equation (5.12),

$$\cos \theta_m \simeq 1 - \theta_m^2/2 = 1 - d/R$$

$$\theta \simeq (2D/r)^{1/2} \tag{5.13}$$

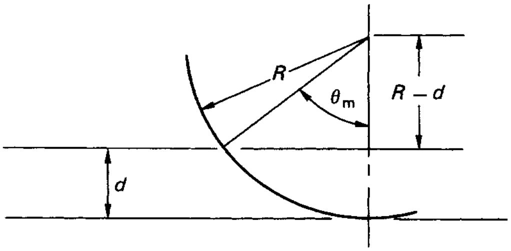


Figure 5.26 Approximate path swept by a cutter

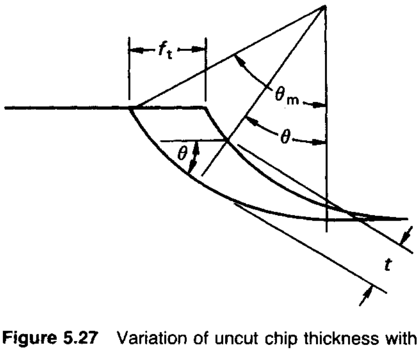


Figure 5.27 Variation of uncut chip thickness with angle of rotation

If the peripheral cutting speed is v m s^{-1} and s is the number of teeth on the cutter, the number of tooth impacts per second will be $1000vs/(2\pi R)$ and, substituting for θ_m, the cutting power with zero helix angle is,

$$P_0 = 2^{(1+x)/2} \, vsw \, Cf_t^x d^{(1+x)/2}/(2\pi(1 + x)R^{(1+x)/2}) \tag{5.14}$$

The volume of metal removed per second is,

$$V = 1000vf_t swd/(2\pi R) \text{ mm}^3$$

Hence, the specific cutting power is

$$P_s = 2^{(1+x)/2} \, CR^{(1-x)/2}/[1000(1 + x)d^{(1-x)/2} f_t^{1-x}] \tag{5.15}$$

Since d is predetermined, it follows that doubling f_t decreases the specific power by 5%, and halving R can reduce it by 5% when cutting steel, assuming $x = 0.85$.

When the cutter has a non-zero helix angle, the quasi-orthogonal model can no longer be applied. The angles by which a helical cutter is specified are the helix angle, which corresponds to the cutting edge inclination λ_s, and the radial rake which corresponds to β' in equation (5.2), from which the normal rake can be calculated.

Cutting power is now modified by including the trigonometric terms in equation (5.11), i.e.

$$P = P_0(\sec^{1-x} \psi_r \cos^x \beta_e \sec^{1+x} \lambda_s \sec^x \gamma_n)$$

For a cylindrical cutter, ψ_r is zero. Thus the formula simplifies to:

$$P = P_0(\cos^x \beta_e \sec^{1+x} \lambda_s \sec^x \gamma_n) = P_0 A \tag{5.16}$$

Consider the three milling cutters of identical size and number of teeth, but with very dissimilar tooth geometry, listed in *Table 5.3*. These cutters are used to machine steel.

The product of the trigonometric terms A for the first two cutters is almost identical, whereas the helical cutter has a product about 60% greater. If the cutting speed is about 1 m s^{-1}, it can be seen from *Figure 5.24* that the cutting force constant C for the second and third cutters is approximately the same, but for the first cutter is about 50% greater. It follows that the cutting efficiencies of the helical and the side and face cutters are roughly similar, but the cutter with the high radial rake is almost 40% more efficient.

Peripheral cutting can be performed in either the upcut or climb mode (*Figure 5.28*). In upcut milling the cutting edge must penetrate the previously cut surface before chip generation commences. This causes a high radial force at the commencement of the cut which does not happen with climb milling. Cutters with large radial rakes have a weak tooth form which results in rapid wear when subjected to the high radial forces associated with upcut milling. Radial forces experienced in climb milling are much lower and wear is usually not a severe problem.

In summary, the most economical performance in terms of specific power can be achieved when climb milling with a high radial rake cutter operated at large feeds. It is preferable to operate at high feeds rather than high cutting speeds since the index of feed in the cutting power equation is less than unity, whereas power is directly proportional to cutting speed.

Table 5.3 Three milling cutters (see text)

	λ_s	β'	γ_n	β_e	A
Side and face cutter	0	10	10.0	10.0	1.0
High radial rake cutter	10	30	29.6	30.6	1.03
Helical cutter	45	10	7.1	34.2	1.63

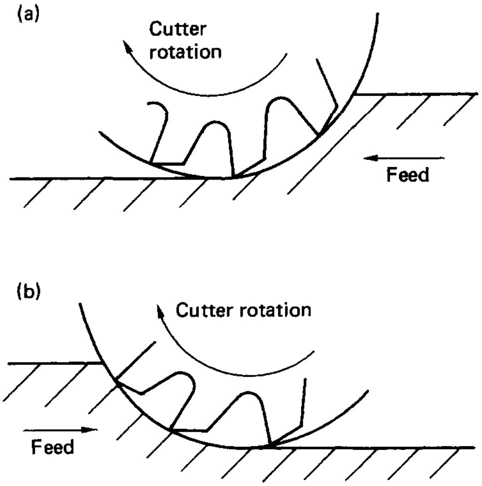


Figure 5.28 Climb and upcut milling: (a) climb; (b) upcut

The foregoing analysis provides useful indicators to the selection of cutting parameters, but it is unlikely that industry would adopt such a complicated method to calculate power requirements. It would be even more difficult to assess the power required for face milling where the work–cutter configuration can vary from total immersion of the cutter face to small arcs of contact in either the upcut or climb modes. A more simple method of power assessment is, therefore, required. This is provided by using the concept of specific energy which is the energy considered necessary to remove unit volume of material. Unfortunately, the influence of cutting speed and rake angle on cutting force, together with the non-linear effect of feed, makes the use of specific energy a very rough guide to cutting power assessment. It is obviously essential to over-predict if a value is to cover all contingencies, so the prediction must be based on very low feeds and negative rakes if it is to be universally applicable to a given work material. The values shown in *Table 5.2* give specific power in watts per cubic millimetre per second, based on the conservative assumption that feeds are of the order of 0.1 mm and that the milling process is about 60% as efficient as turning.

5.5.3 Hot machining

Some work materials pose machining problems which cannot readily be solved by conventional methods. These include the nimonic alloys and Stellite cast alloys. For example, some of the nimonic alloys when machined with carbide tools on a 100 mm diameter bar, necessitate cutting speeds as low as 0.15 m s^{-1} and tool failure commonly occurs after machining a 100 mm length of the bar. When the surface of the bar is preheated with a gas tungsten arc or a transferred plasma arc struck between the electrode and the work surface and using ceramic or cubic boron nitride tools, cutting speeds of about 2.5 m s^{-1} are possible, and the tools remain serviceable after machining a considerable length of bar.

This technique is not one which would be advised if alternatives are possible but, with some of the more refractory metals now in use, the hot machining process is frequently the only practical solution. The surface preheat temperature is about 600°C, giving such high interface temperatures that carbide tools cannot be used.

5.6 Surface-finish considerations

Built-up edge is one of the main factors contributing to poor surface finish. When machining most materials this can be reduced, if not eliminated, by operating at high cutting speeds. Where finish cuts are required, the uncut chip area is relatively small, so the cutting power is never likely to be an important consideration, even at very high cutting speeds. If ceramic tools are used the speed limitation is usually that imposed by the available spindle speeds, but when using carbides or high speed steels the speed constraint is usually that imposed by tool wear.

The theoretical surface roughness in turning is determined by tool plan geometry, a pointed tool operated at a given feed producing a rougher surface than one having a nose radius (*Figure 5.29*). In Britain, the most popular measure of surface roughness is the centre line average, which is the average deviation from the mean line drawn through the surface. For a pointed tool the theoretical surface roughness is given by

$$R_a = f/(4 \tan \psi_r + \cot \kappa_{r'}) \qquad (5.17)$$

and for a tool of nose radius r

$$R_a = f^2/(31.2r) \qquad (5.18)$$

In equation (5.18) the surface roughness is proportional to the square of the feed and inversely proportional to the nose radius. A large nose radius is obviously desirable but, since this is likely to induce chatter, most disposable tips are limited to a radius of about 1 mm.

Due to imperfect chip formation and surface roughness of the tool, the theoretical value is never achieved in practice. Rake angle and width of cut have very little effect on the value obtained. *Figure 5.30* shows how surface roughness varies with cutting speed and compares it with the theoretical value for a tool having a feed of 0.075 mm and a nose radius of 1 mm.

The surface generated by a peripheral milling cutter is directional in property. In the direction of feed the theoretical surface is geometrically similar to that for a turned surface, the cusps having a radius equal to that of the cutter, and the pitch between cusps being equal to the feed per cutting tooth. Due to the almost inevitable lack of straightness of the arbor on a horizontal milling machine, the contour generated by the cutter teeth varies as some teeth take a greater depth of cut than others. In severe cases, one tooth may take such a disproportionately deep cut that the surface generated has a periodicity corresponding to the feed per revolution rather than the feed per tooth, and the cusps are correspondingly deeper.

Face milling usually produces a finish superior to that generated by peripheral cutters. The geometry of the cutter is specially designed so that the combination of corner angle, end cutting edge angle and nose radius produce very flat cusps. In addition, due to the use of carbide cutters, cutting speeds are much higher than those achievable using high speed steel peripheral cutters.

5.7 Tool-life assessment

It is fortunate that, with few exceptions, tool wear occurs in a predictable manner. Although it takes different forms, each form of tool wear is associated with a known cause which happens within particular ranges of feed and cutting speed. The production engineer's initial task in optimising the metal cutting process is to decide which form of wear is to be accepted, and then to work within the constraints which such wear imposes. Before pursuing this line of thought further, it

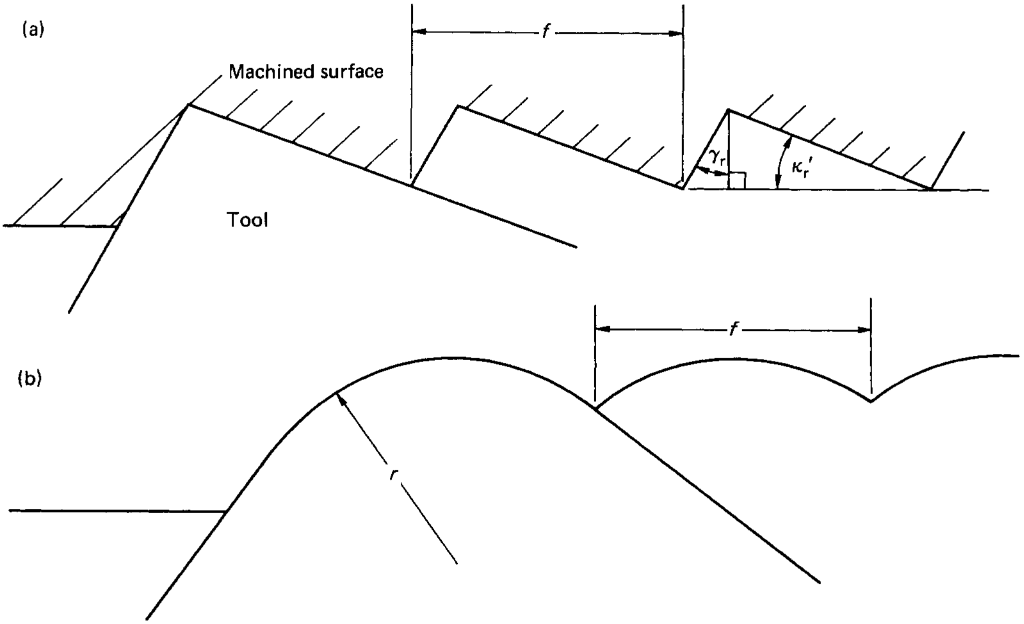


Figure 5.29 Surfaces generated by single-point tools: (a) pointed tool; (b) round-nosed tool

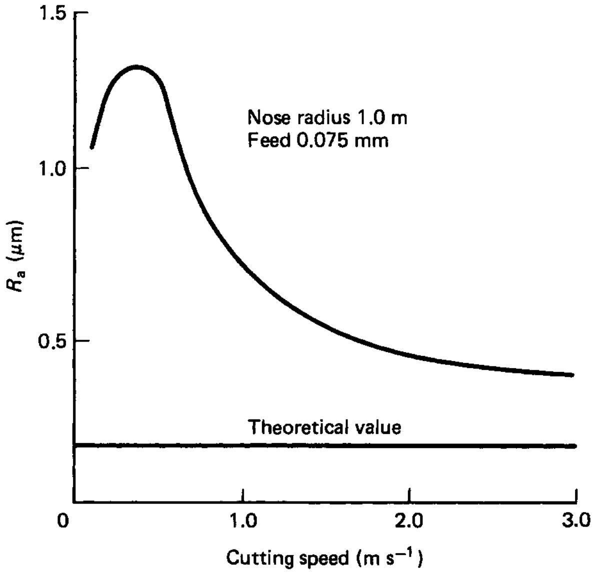


Figure 5.30 Comparison of experimental and theoretical values of surface roughness

is desirable to consider the various wear patterns which can occur.

5.7.1 Crumbling of the cutting edge

As mentioned earlier, when cutting steel at low speeds, built-up edge is a persistent problem. The build up is initiated by purely mechanical causes but, if it is allowed to remain *in situ* for sufficient time, diffusion welding occurs between the tool and the stationary built-up edge. If a carbide tool is used, the coefficient of expansion is less than that for a steel tool, so a subsequent drop in temperature, as for instance when the tool is taken out of cut, causes tensile stresses in the surface of the tool which can result in flakes of carbide being pulled from the surface. After this has been repeated a few times, the cutting edge will have deteriorated, giving the appearance of crumbling.

The solution when machining steel is fairly straightforward. An increase in cutting speed, if necessary at the expense of feed, will reduce the problem of build up and prevent crumbling.

5.7.2 Cratering

When tungsten carbide was used as the main component of sintered carbide tools, it was common for a crater to form near the hottest portions of the rake face due to diffusion between the tool and the chip, tool material being carried away on the underside of the chip. Since the high temperatures were largely the result of high cutting speeds, and to a lesser extent large feeds, the obvious solution was to restrict the cutting speed.

This problem has been reduced considerably by applying a surface coating of titanium carbide to the tool which diffuses less readily. Crater wear still occurs, but it is seldom the main wear source and judicious selection of cutting speed and feed can usually be made to reduce its effect.

5.7.3 Plastic deformation of the tool

Most materials subjected to high stresses and high temperatures start to creep. On a cutting tool this can result in the cutting edge deforming, causing the tool geometry to change

and producing high cutting forces. These forces, in turn, raise the temperature and cause greater deformation to occur.

If the interface temperature and stress levels are reduced, creep can be prevented. A surface coating which reduces friction, such as titanium nitride, reduces the rise in interface temperature and reduces plastic deformation. Usually, unless the cutting speed and feed are very high, creep is not a severe problem.

5.7.4 Thermal and mechanical shock

Although these are different phenomena, they usually occur in combination. Mechanical shock occurs with a sudden change in chip formation, such as a turning tool crossing a keyway or a milling tool coming into or leaving engagement. Inevitably, discontinuities of this sort also cause thermal shock.

Mechanical shock, inducing tensile stresses, can cause small particles to break away from the cutting edge. Thermal shock causes a temperature gradient to occur along the cutting edge. Tensile stresses resulting from this can cause hair-line cracks which act as stress concentrations and result in breakdown of the cutting edge.

A main contributor to thermal shock is the application of inadequate coolant. For this reason it is usually advisable when using brittle tool materials such as carbides or ceramics to cut dry. In general, the limitation of mechanical shock may be relaxed by reducing feed and, to a lesser extent, cutting speed.

5.7.5 Flank wear

When all other forms of wear are eliminated or substantially reduced, tools will exhibit a progressive form of attritive wear, causing a scar on the front clearance face or flank. This is the mechanism by which most tools fail, and a study of metal-cutting economics must start from a consideration of what constitutes a reasonable amount of wear before replacing the cutting edge.

5.7.6 Catastrophic failure

Although cutting tools have been known to fail catastrophically for no obvious reason, such failures in modern carbide tools are so rare that they can be ignored. Generally, when catastrophic failure occurs it is a direct consequence of excessive wear due to one of the other modes of failure. If the tools are regularly inspected for wear they can almost invariably be withdrawn from service before catastrophic failure becomes a significant probability.

5.7.7 Tool-life relationships

The ranges of feed and speed within which the various wear mechanisms predominate will vary considerably with different tool/work material combinations.[4] *Figure 5.31* shows a typical set of boundaries within which cratering, plastic deformation, crumbling of the cutting edge or flank wear predominate. These boundaries are usually so well separated that the area due to flank wear provides a wide range of both feeds and cutting speeds from which the cutting parameters can be chosen, these ranges being shown dotted in the figure.

If cutting tests are conducted under constant conditions, the wear scar will be seen to increase in three fairly well-defined stages. In the first stage wear is fairly rapid as the tool loses its initial sharpness. This is followed by a secondary stage of slow, almost constant wear rate. The third stage, which commences when the wear scar approaches 1 mm in length, is due to rapidly rising temperature at the cutting edge caused by secondary heating on the rake face together with a third heat source due to rubbing on the flank face.

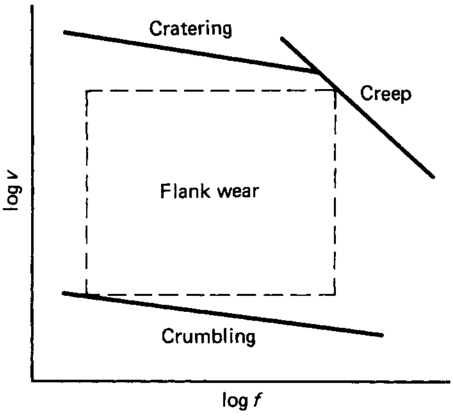


Figure 5.31 Tool-wear envelope

When tests are conducted at different cutting speeds, a family of curves similar to the one shown in *Figure 5.32* will be produced. The third stage develops rapidly and, if allowed to continue, will result in catastrophic failure. It is therefore desirable to specify a wear scar near the boundary between the second and third stages of wear. This usually occurs with a flank wear scar of about 1.0 mm when using high speed steel tools and 0.75 mm when using carbides.

If the time to produce scars of this magnitude is plotted against cutting speed on logarithmic scales, a straight line relationship results (*Figure 5.33*), giving rise to a tool-life equation of the form[5]

$$vT^n = \text{constant} \tag{5.19}$$

It is fortunate that the exponent n varies very little across the whole range of steels, taking values of about 0.13 for high speed steel tools, 0.20 for negative rake carbides and 0.25 for positive rake carbides. Since most high production cutting is performed with carbides, the second and third values are of greatest importance. Values of about 0.38 have been quoted for ceramic tools, but these appear to be based on scanty information and are probably unreliable.

Equation (5.19) is important in that it shows the influence of cutting speed on tool life. It is very restrictive in that it does

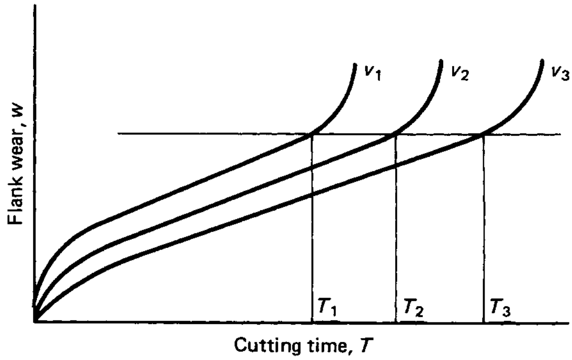


Figure 5.32 Flank wear related to cutting time

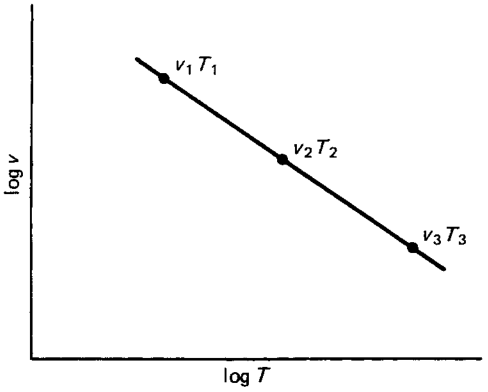


Figure 5.33 Cutting speed vs. tool life relationship

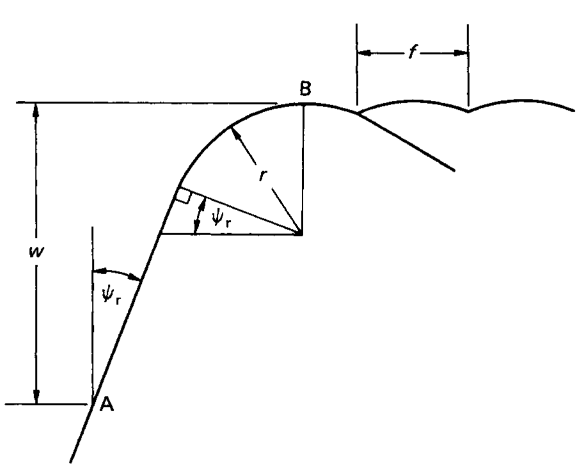


Figure 5.34 Tool plan geometry

not allow for variation in feed, width of cut or tool geometry. Well documented evidence exists which enables this equation to be extended to embrace feed variation, as follows,

$$vT^nf^m = \text{constant} \tag{5.20}$$

Again, it is fortunate that the exponent m does not appear to vary appreciably when cutting any of the steels and is unaffected by rake angle. A value of 0.37 can be assumed.

The next stage in generalising the equation is far more tentative, being based on intuition rather than experimental data. Several researchers have suggested a somewhat similar approach, but the main significance of the generalised equation can be attributed to Brewer.[6] He reasoned that, since tool wear is a function of interface temperature, a temperature equilibrium must exist which is determined by the ratio of heat generation and heat conduction from the cutting edge. The rate of heat generation is assumed to be proportional to the uncut chip area (fw) and the rate of heat dissipation is assumed to be proportional to the length of cutting edge in engagement.

If a tool taking a width of cut w and a feed f has an approach angle ψ_r and a nose radius r the effective length of the cutting edge can be shown to be approximately equal to the line AB in *Figure 5.34*, i.e.

$$l_e = [w - r(1 - \sin\psi_r)]/\cos\psi_r + (90 - \psi_r)\pi r/180$$

and the interface temperature θ is a function of fw/l_e. This ratio has the dimensions of length and is referred to as the 'equivalent chip thickness', b_e. It includes the cutting variables f, w, ψ_r and r, of which f is readily separable,

$$b_e = fw/l_e = fG$$

G has a maximum value of unity for a pointed tool with zero tool approach angle, and in practical situations is usually about 0.9.

A generalised equation including these variables could, therefore, be of the form

$$vT^nb_e^m = vT^n(fG)^m = \lambda \tag{5.21}$$

where λ is constant for any given work and tool material and

$$G = w/[(w - r(1 - \sin\psi_r))/\cos\psi_r + (90 - \psi_r)\pi r/180] \tag{5.22}$$

Purists may argue that the assumptions underlying this equation are so sweeping as to make it unrealistic. Its intuitive appeal lies in the fact that a longer cutting edge due to a non-zero tool approach angle and a nose radius would be expected to reduce the effective value of uncut chip thickness, which is what the value of G achieves.

5.8 Economics of metal cutting

The main financial objective of a manufacturing company is to maximise the return on capital expenditure. This implies a knowledge of profitability, which in turn requires a knowledge of cost and selling price. Production engineers are concerned with the processes needed to make components which are eventually assembled into finished products. Selling price, therefore, is not usually a very useful statistic for the process planner. He/she must settle for subobjectives such as minimum cost or, sometimes, maximum output which, although not synonymous with profitability, at least contribute to its achievement. When discussing manufacturing economics, these are the objectives which must be addressed.

Metal cutting is an intrinsically wasteful operation, involving the removal of large quantities of material. Although there are no reliable figures to support this contention, it is probable that only about 70% of the material purchased is contained in finished parts, the balance being expensively converted to swarf which has a very low resale value. Intelligent design can do much to increase material utilisation, but material wastage will always be a significant proportion of the total component cost. In spite of the attractiveness of contending production options such as metal forming, it is inevitable that cutting processes will continue to be used extensively. The subsequent analysis assumes that due cognisance has been taken at the design stage of the importance of material utilisation, and the cost factors include only the direct cost of manufacture and its associated overhead. The operating cost, taking into account direct labour, machine depreciation and factory overheads, may well be of the order of £20 per hour.

5.8.1 Minimum-cost production

The cost of manufacture can be divided into five parts:

(1) set-up and idle time cost per component, K_1;
(2) machining cost per component, K_2;
(3) tool changing cost per component, K_3;

(4) tool depreciation cost per component, K_4; and
(5) tool regrinding cost per component, K_5.

Assuming that disposable inserts are used, there is no regrinding cost, so this can be ignored.

The setting cost can be substantially reduced by using preset tooling. However, with the current trend towards small batch sizes, the setting cost ascribed to each component will increase proportionally. The idle time per cutting cycle is composed of loading and unloading time in addition to the tool approach and tool retraction times before and after machining has taken place. Set-up and idle time can, therefore, contribute significantly to production cost, and its reduction is frequently the largest single factor in cost minimisation.

Machining cost is directly related to cutting time which, in turn, is dependent on cutting speed and the size of feed. The use of large feeds and high cutting speeds reduces machining cost but reduces tool life and, consequently, increases both the unit tool depreciation and unit tool changing costs. It is therefore the minimisation of the total of these three costs which determines the minimum production cost for any given set-up.

The simplest analysis, and in fact the only analysis which has been exhaustively attempted, is that for a plain bar turning operation.[7] Fortunately, the conclusions to be drawn from the bar turning analysis provide good indicators when seeking to optimise other turning and milling operations.

If a metal bar of diameter D mm is machined along a length L mm with a cutting speed v m s^{-1} and a feed f mm, the machining time,

$$t_m = L\pi D/(60\,000fv) \text{ min}$$

from equation (5.21), the tool life is

$$T = \lambda^{1/n}/[v^{1/n}(fG)^{m/n}]$$

and the number of tool changes per component is

$$t_m/T = L\pi Dv^{(1/n)-1}f^{(m/n)-1}G^{m/n}/(60\,000\lambda^{1/n})$$

The set-up and idle time cost per component is

$$K_1 = k_1t_s$$

where k_1 is the cost per minute of direct labour, overhead and machine tool depreciation, and t_s is the set-up and idle time per component.

The machining cost per component is

$$K_2 = k_1L\pi D/(60\,000fv)$$

The tool changing cost per component is

$$K_3 = k_1t_cL\pi Dv^{(1/n)-1}f^{(m/n)-1}G^{m/n}/(60\,000\lambda^{1/n})$$

where t_c is the time needed to change a tool.

The tool depreciation cost per component is

$$K_4 = EL\pi Dv^{(1/n)-1}f^{(m/n)-1}G^{m/n}/(60\,000\lambda^{1/n})$$

where E is the tool cost per cutting edge.

The total manufacturing cost per component is

$$K = K_1 + K_2 + K_3 + K_4$$
$$= k_1t_s + k_1L\pi D/(60\,000fv) + L\pi Dv^{(1/n)-1}f^{(m/n)-1}G^{m/n}(k_1t_c + E)/(60\,000\lambda^{1/n}) \quad (5.23)$$

The three variables in this equation over which the production engineer has some control are f, v and G. By inspection it is seen that a low value of G is desirable. This can be achieved by having a large tool approach angle and a large nose radius. In practice, these may be limited by geometrical requirements of the work piece or by the likelihood of chatter, and they are usually limited to 15° and 1.0 mm, respectively, when using disposable tips.

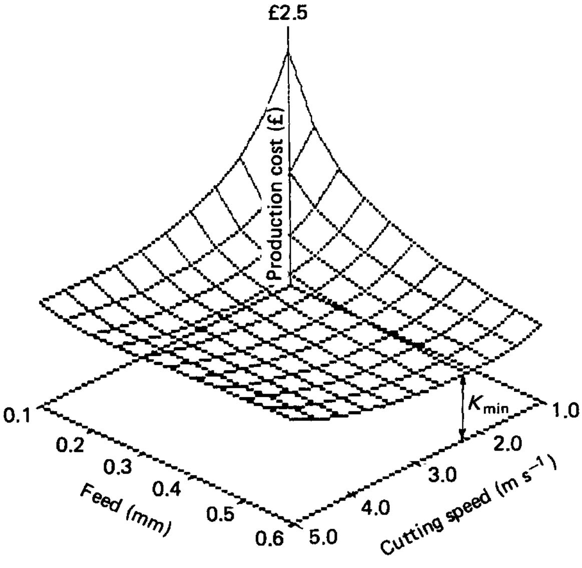


Figure 5.35 Production cost related to cutting speed and feed

Figure 5.35 illustrates by way of a carpet plot how K varies with both f and v. The main point worth noting is that, within the broad range of feeds and cutting speeds selected, the minimum cost occurs inside the speed range and at the maximum value of feed.

It would be useful to possess a set of values for λ covering all work materials and up to date cutting materials, but the few values which exist relate to carbide tools which were commercially available several years ago. Thus, *Figure 5.35* relates to a test using positive rake carbide tools to cut EN24 steel bar having a Brinell hardness of 280, for which the value of λ is 3.5. Using currently available coated tools this value would be expected to be somewhat larger. The other variables were given the following values:

Tool cost per cutting edge E = £1.0
Diameter of bar D = 100 mm
Length of cut L = 150 mm
Machine rate per minute k_1 = £0.25
Set-up and idle time per component t_s = 2.0 min
Tool changing time t_c = 1.0 min

This graphical model was used to predict the effects on manufacturing cost of varying the parameters singly and in combination. The most significant observations obtained from these simulations, together with the graphs obtained are detailed below.

(1) The value of G has a very small effect on the optimal cutting speed or on minimum cost across the practical range of values for G of 1.0–0.8. This would seem to suggest that the effects of tool plan geometry and of width of cut are so slight that they can be ignored for practical purposes. *Figure 5.36* shows cost graphs obtained using the two extreme values of G.
(2) Minimum cost is almost directly proportional to the machine rate k_1 (see *Figure 5.37*). Since this parameter comprises the sum of depreciation rate, direct labour cost

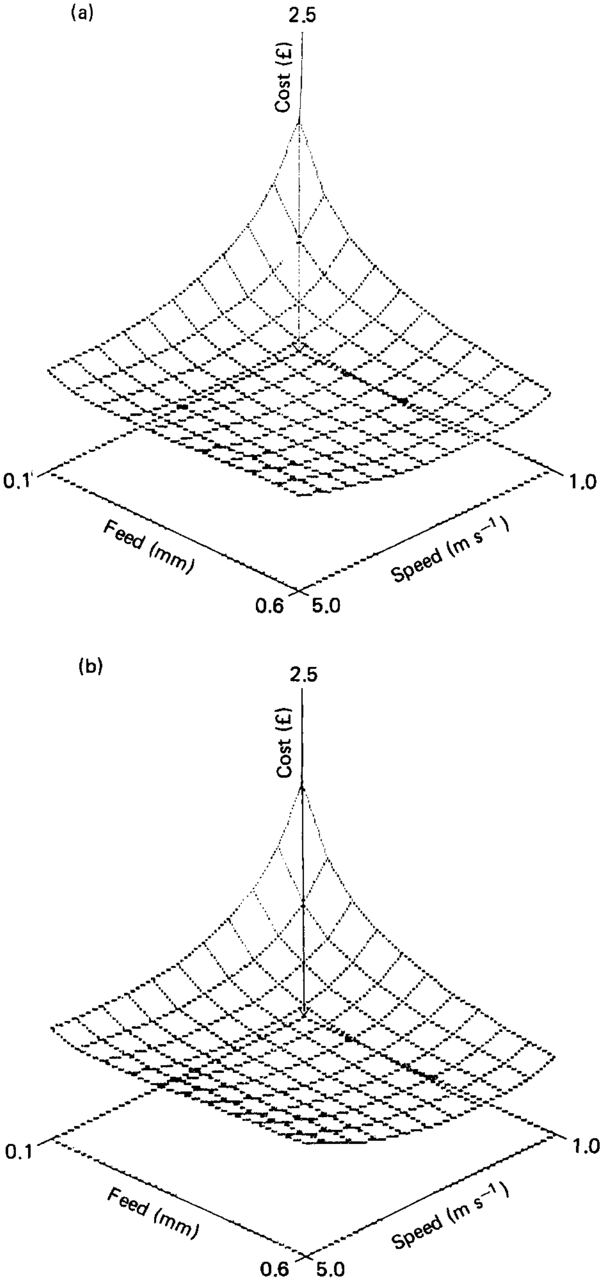


Figure 5.36 Effect on cost of varying G: (a) G = 0.8; (b) G = 1.0

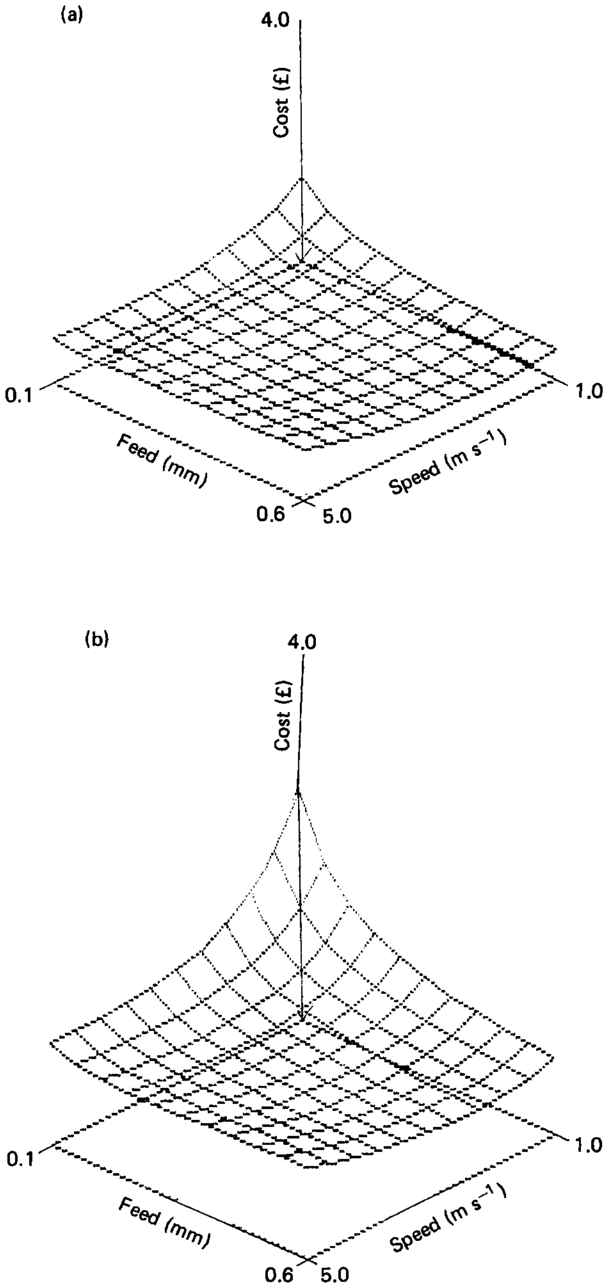


Figure 5.37 Influence of machine rate per minute on cost: (a) k_1 = £0.15; (b) k_1 = £0.40

and factory overhead it is obvious that, when using expensive equipment, round-the-clock operation and high utilisation are very important.

(3) The value of λ has a dramatic effect on the shape of the cost curve and on the minimum cost values, as shown in *Figure 5.38*. Low values of λ give a very cost-sensitive curve, with low optimal cutting speeds. Large values of λ, associated with the newer high performance cutting materials, produce curves which are very flat and insensitive to cost near the minima, which occur at high cutting speeds. High values therefore have the double advantage of low costs and of considerable flexibility in selecting cutting speeds, which makes for high productivity with little penalty in cost.

(4) Tool-changing time has relatively little effect on minimum cost but, as shown in *Figure 5.39*, long changing times cause a greater cost sensitivity to variations in cutting speed.

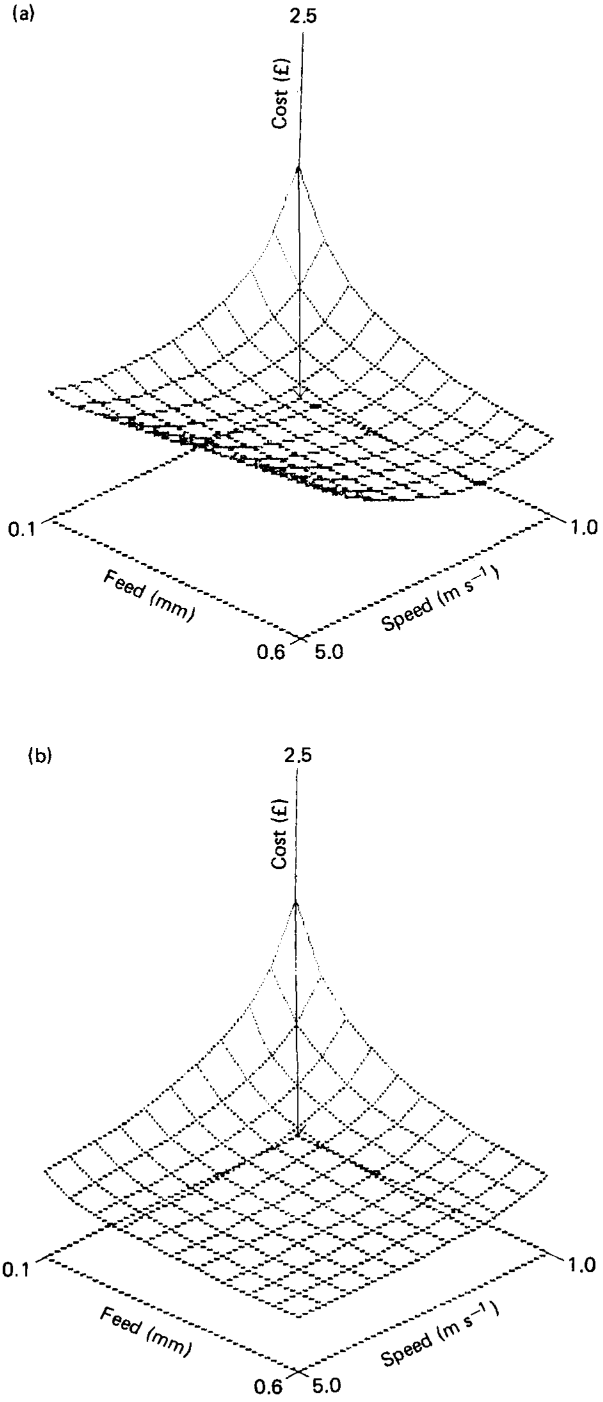


Figure 5.38 Variation of cost due to varying λ: (a) λ = 3.0; (b) λ = 7.0

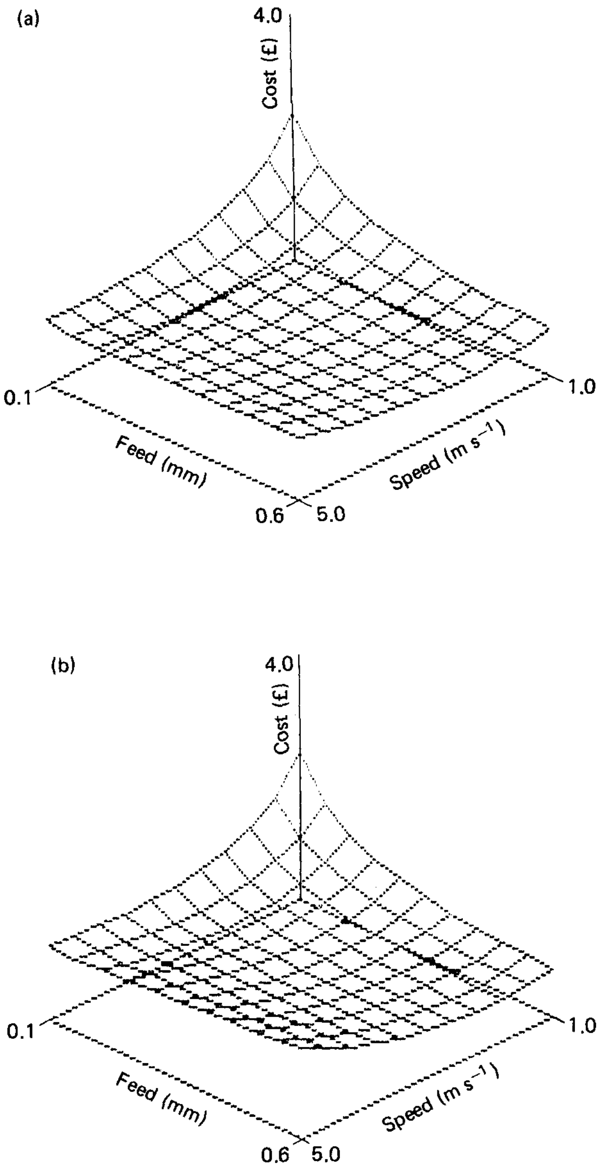


Figure 5.39 Effect of tool changing time on cost: (a) t_c = 0.4 min; (b) t_c = 4.0 min

(5) Total idle time has marked effect on minimum cost (*Figure 5.40*). This comprises setting time attributed to each unit of output, and to the non-cutting time in each manufacturing cycle (due to loading and unloading, and also to tool approach and retraction time). A strong case exists for streamlining batch changeovers, for improving materials handling, and for more rapid approach and withdrawal speeds.

Although reliable values of λ are not available from which to calculate optimal cutting speed, it is possible, as shown below, to establish the optimal tool life. Knowing the maximum feed which can be selected the appropriate cutting speed can be arrived at by experiment, i.e. by varying the speed to

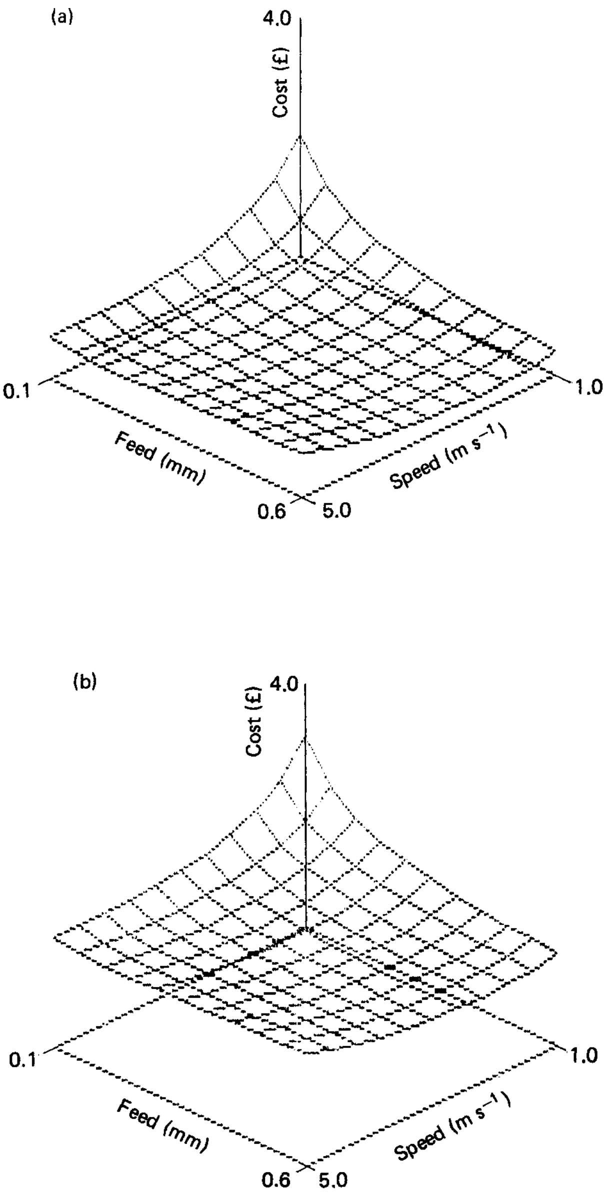


Figure 5.40 Influence of idle time on cost: (a) t_s = 0.4 min; (b) t_s = 5.0 min

give the desired tool life. The maximum practical feed may be determined by the largest available on the lathe, by the maximum permissible cutting force, or by surface roughness considerations.

Differentiating equation (5.23) with respect to cutting speed,

$$dK/dv = -k_1 L\pi D/(60\ 000 f v^2) + (1/n - 1)$$
$$L\pi D v^{(1/n)-2} f^{(m/n)-1} G^{m/n}(k_1 t_c + E)/(60\ 000\lambda^{1/n})$$

Equating to zero for minimum-cost conditions:

$$k_1 = (1/n - 1)v^{1/n}(fG)^{m/n}(k_1 t_c + E)/\lambda^{1/n} \quad (5.24)$$

From equation (5.21),

$$T = \lambda^{1/n}/(v^{1/n}(fG)^{m/n})$$
$$T_{opt} = (1/n - 1)(k_1 t_c + E)/k_1 \quad (5.25)$$

It is clear that T_{opt} could vary considerably depending on the values of k_1, t_c and E. Taking the values used in *Figure 5.35*, i.e. k_1 = £0.25, t_c = 1.0 min and E = £1.0 per cutting edge,

$$(k_1 t_c + E)/k_1 = 5.0$$

When cutting steel with positive rake tools, n = 0.25, giving an optimum tool life of 15 min. Even if the cost term in equation (5.25) increased by a factor of 2 the optimal tool life would still be only 30 min, which is considerably less than that normally used in industry.

This presupposes that sufficient power is available to permit the use of feeds and cutting speeds high enough to cause the tools to wear out in the optimal time. Consider a fairly high powered lathe with an available spindle power of 10 kW, capable of operating at large feeds and high spindle speeds, taking a 5.0 mm width of cut on EN24 steel, for which the cutting force constant C is 2000 when using positive rake carbide tools:

Power P = Principal cutting force × Cutting speed

$$P_{max} = 10\ 000 = 2000 \times 5f^{0.85}v_p$$
$$v_p = 1/f^{0.85}$$

where v_p is the maximum speed constrained by spindle power. Assuming the values taken to produce *Figure 5.35*,

$$v_m = 1.849/f^{0.37}$$

where v_m is the cutting speed giving minimum cost.

Figure 5.41 shows that for any feeds in excess of 0.3 mm the available power constrains the maximum cutting speed to a value less than that which would produce a 15 min tool life. In most practical situations of heavy metal removal the power constraint is even more severe, so the case for operating at maximum power assumes added importance.

What conclusions can be drawn from this analysis when considering the other main heavy metal removal process, face milling? Obviously, with a multipoint cutter the optimal tool life would be greater than for a single-point cutter. Although

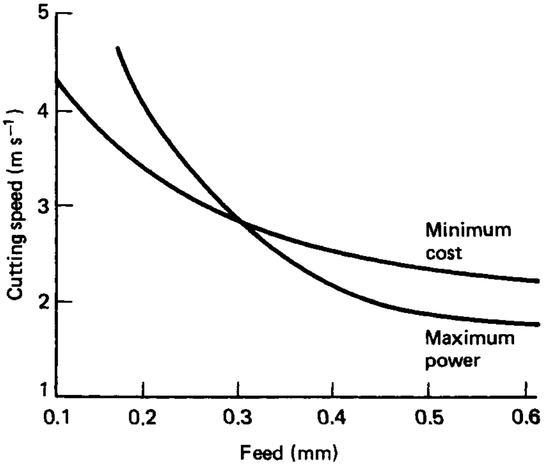


Figure 5.41 Typical minimum cost and maximum power constraints

no comparable analysis exists, it would seem reasonable that, if the optimum life for a single-point cutter is 15 min, that for an eight-tooth cutter would be of the order of 2–3 h.

With multipoint cutters the problems associated with catastrophic failure are less acute, since the failure of one cutting edge will not pose a dangerous threat. Usually, cutting is allowed to proceed until the generated surface is deemed to be unsatisfactory. This happens after incurring more severe wear than would be permissible with a lathe tool. To approach an optimal solution to the face milling process it would appear, therefore, that the maximum recommended feed per cutter tooth should be selected and the spindle speed should be set to provide a cutter life of about 3 h if this is possible. More likely, the power constraint will override the economic constraint, in which case the cutter should be run at a spindle speed approaching that requiring maximum available power.

5.8.2 Machining for maximum output

Whereas minimum-cost machining is based on optimising the equation for unit cost, maximum productivity is achieved by minimising the unit manufacturing time.

The time taken to produce one component is given by

$$M = t_s + t_m + t_c \times \text{Number of tool changes per component}$$
$$= t_s + L\pi D/(fv) + t_c L\pi D v^{1/n}(fG)^{m/n}/(f v \lambda^{1/n})$$

Differentiating with respect to cutting speed, and equating to zero:

$$v_0 = \lambda/[(1/n - 1)^n (fG)^m t_c^{\,n}] \quad (5.26)$$

For minimum cost, from equation (5.24),

$$v_m = k_1^n \lambda/[(1/n - 1)^n (fG)^m (k_1 t_c + E)^n]$$

Since $1/t_c > k_1/(k_1 t_c + E)$, it follows that, for a given maximum feed, the cutting speed giving maximum output v_0 will always be greater than that giving minimum cost v_m. This would demand very short tool lives, having disproportionately large standard deviations, and a high probability of catastrophic failure. Even if the drive motor possessed sufficient power to operate under these conditions it is not normally to be advised. The sole purpose in extending the analysis to encompass maximum output is to show that the short tool lives associated with minimum-cost machining assist also in improving productivity.

5.8.3 The influence of chatter

A great deal of research has been published on the stability boundary between width of cut and spindle speed.[8] The boundary envelopes vary with the machine tool and the process parameters, but a typical boundary is shown in *Figure 5.42*, where the maximum width of cut under stable machining conditions decreases as the spindle speed increases. A large tool approach angle and nose radius increase the effective length of the cutting edge which is tantamount to increasing the width of cut and decreasing the uncut chip thickness. If chatter is likely to occur it is obvious that tool approach angle and nose radius should be kept as small as possible, a recommendation which conflicts with recommendations as regards tool life.

The influence of feed on the stability boundary is more significant, but this does not appear to have received the attention it deserves. *Figure 5.43* shows how the critical width of cut increases as feed increases, making the case for using large feeds to oppose the onset of chatter.[9]

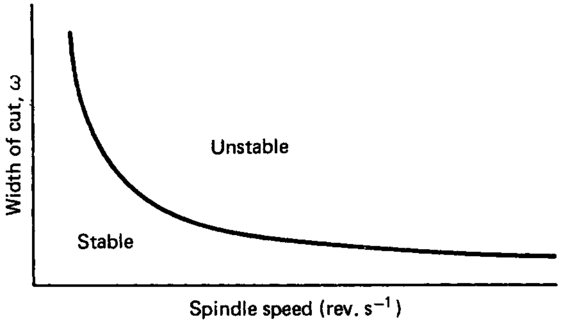


Figure 5.42 Variation of stability threshold with spindle speed

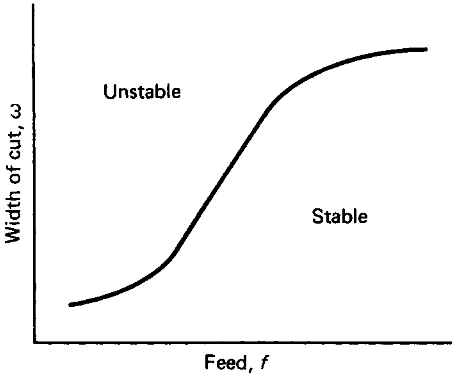


Figure 5.43 Variation of stability threshold with feed

In conclusion, it is pertinent to stress again the importance of large feeds in achieving not only chatter-free performance, but also low specific cutting power and minimum-cost machining.

References

1 BRITISH STANDARDS INSTITUTION, *BS 1296 Specification of single point cutting tools*, Milton Keynes (1972)
2 STABLER, G. V., 'The fundamental geometry of cutting tools', *Proc. Inst. Mech. Eng.*, **165**, 14 (1951)
3 CROOKALL, J. R. and RICHARDSON, D. B., 'Use of photographed orthogonal grids and mechanical quick stopping techniques in machining research', *Conference on Photography in Engineering*, Institute of Mechanical Engineers, London (1969)
4 TRENT, E. M., *Metal Cutting*, Butterworth, London (1977)
5 TAYLOR, F. W., 'On the art of cutting metals', *Trans. ASME*, **28**, 31 (1907)
6 BREWER, R. C., 'On the economics of the basic turning operation', *Trans. ASME*, **80**, 1479 (1958)
7 BREWER, R. C. and RUEDA, R. A., 'Simplified approach to the optimum selection of machining parameters', *Eng. Digest*, **24**, 9 (Sep. 1963)
8 TOBIAS, S. A., *Machine Tool Vibrations*, Blackie, London (1965)
9 PEARCE, D. F. and RICHARDSON, D. B., 'Improved stability in metal cutting by control of feed and tool/chip contact length', *Joint Polytechnic Symposium on Manufacturing Engineering*, Leicester (1977)

6

Non-chip Metal Removal

D Koshal

Contents

6.1 Introduction 6/3

6.2 Electrical processes 6/3
- 6.2.1 Electrical discharge machining 6/3
- 6.2.2 Electrochemical machining 6/4
- 6.2.3 Electrolytic grinding 6/4
- 6.2.4 Other electrochemical processes 6/5

6.3 Mechanical processes 6/5
- 6.3.1 Water-jet machining 6/5
- 6.3.2 Ultrasonic machining 6/5

6.4 Thermal processes 6/7
- 6.4.1 Electron-beam machining 6/7
- 6.4.2 Laser-beam machining 6/7
- 6.4.3 Plasma-arc machining 6/8

6.1 Introduction

There are applications where metal removal by chip formation is not satisfactory or economical for the following reasons:

(1) the tensile strength and hardness of the material is very high;
(2) the shape of the part is too complex;
(3) better surface finish than that possible with chip removal processes is required;
(4) temperature gradient or residual stresses in the work piece are undesirable; and
(5) miniaturisation in the electronic industry.

These requirements led to the development of a number of non-traditional machining processes. The most commonly used processes are described in this chapter.

6.2 Electrical processes

6.2.1 Electrical discharge machining

This process is sometimes referred to as 'electro-erosion' or 'spark machining'. It may be described briefly as the removal of material by means of repetitive short-lived electric sparks which occur between the tool (i.e. electrode) and the work piece.

The principle of the process is shown in *Figure 6.1*, in which the tool is mounted on a chuck or other clamping device attached to the spindle of the spark machine.The down feed of this spindle is controlled by a servodrive through a reduction gearbox. The work piece (i.e. cathode) is immersed in a tank filled with a dielectric fluid, usually paraffin; in order to reduce the risk of fire a depth of several inches is maintained over the top of the work piece. A dielectric fluid is circulated under pressure by a pump and filtered to remove vaporised work-piece particles.

The circuit shown in *Figure 6.1* is a d.c. relaxation circuit fed usually from a mercury arc or selenium type rectifier. In order to achieve erosion, a gap (approximately 0.025–0.05 mm) between the tool and the work piece has to be maintained by the servodrive. The frequency of spark production is of the order of 10 000 per second. As the current is discharged across the gap, temperatures in the region of 10 000–50 000°C are developed, this evaporates parts of the surrounding dielectric fluid and vaporises the metal, thus forming a small crater on the work surface.

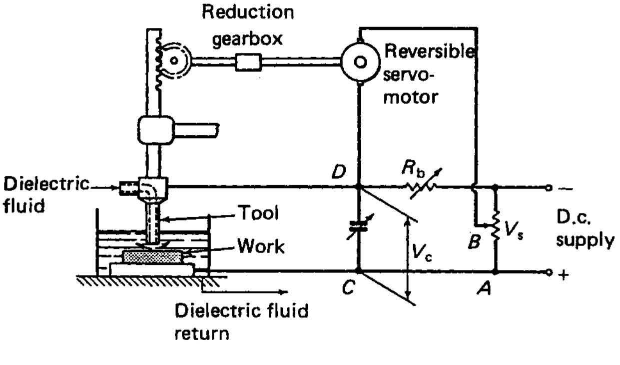


Figure 6.1 Electrical discharge machining apparatus

6.2.1.1 *Tool wear and tool materials*

The fact that wear occurs upon the tool (i.e. electrode) is one of the major difficulties in electrical discharge machining (EDM). The wear ratio, i.e. the metal lost from the tool divided by the volume of metal removed from the work, will depend on the tool and work materials used. Typical examples are that of a brass tool used on a brass work piece, where the ratio is of the order of 0.5, whereas for a brass tool used on a tungsten carbide work piece the ratio may be as high as 3.

Electrode wear will produce inaccurate machining, the wear taking place predominantly at the leading edges of the electrode, thus producing tapered cavities. Advantage may be taken of this feature when producing press tool blanking dies, the electrode being put through the die from the reverse side, thus providing natural die clearance. In most cases, however, electrode wear is not an advantage as it may be necessary to use several electrodes in order to achieve acceptable accuracy.

In order to minimise the cost of electrode manufacture, much work has been done to obtain suitable materials for the electrodes. For many applications brass or copper is used, although copper–graphite and tungsten carbide have been used to advantage. The higher the melting point of the electrode material the less wear occurs. Thus, by using materials such as copper–graphite and, more recently, with the pulse generator circuit, graphite, lower wear rates are obtained.

6.2.1.2 *Metal removal*

If the highest quality surface finish is required, then lower metal removal rates must be used. Thus it will be necessary to achieve a compromise, depending on whether rough or finish machining is required.

Using modern pulse generator machines, metal removal rates of up to 8 mm^3/min can be achieved when rough machining. The metal removal rate is affected by the spark duration for a given energy input. For a given power input an optimum discharge time exists, thus giving maximum metal removal per discharge. If the discharge time is too short, tool wear becomes excessive with a resulting loss of accuracy.

6.2.1.3 *Accuracy*

In addition to errors occurring due to incorrect discharge times, the accuracy with which the electrode is produced is clearly an important factor. Allowance must be made for the spark gap, and parallel-sided cavities can be produced by using several electrodes or by relieving the tool behind the tip. As the depth of the cavity increases it becomes more difficult for the dielectric fluid to wash the disintegrated work particles clear of the tool–work gap. This has the effect of reducing the breakdown voltage of the dielectric fluid with the result that sparking occurs between the sides of the electrode and the work cavity.

6.2.1.4 *Surface roughness*

The surface finish in EDM depends on the rate of metal removal. Good finish requires low power discharges that leave small craters, but the rate of metal removal is slow. It is generally advisable to use roughing electrodes first and then finish off. When high production rates are required, the surface roughness R_a may be as high as 1.75–3.75 μm, but by reducing the spark rate R_a values in the region of 0.25 μm may be achieved.

6.2.1.5 Applications

The application of EDM is limited to the machining of electrically conducting work piece materials, but the process has the capability of cutting these materials, regardless of their hardness or toughness.

Spark machining has its greatest application in the manufacture of press tools, mould forgings and extrusion dies. The material can be spark machined after it has been hardened, thus eliminating the effects of distortion due to heat treatment. Tools manufactured from cemented carbide can be machined when sintering has been completed, thereby eliminating the need for an intermediate partial sintering stage and thus eliminating the errors resulting from final sintering after the cavities, etc., have been machined.

The successful application of this method to the drilling of holes with a large length/diameter ratio, as in the case of the nozzles of diesel engine fuel injectors, has been carried out.

In addition to the conventional applications of EDM, the process has been used for turning, grinding, and precision band sawing. In the former case the work piece is rotated as in turning, but the cutting tool is replaced by the electrode. The dielectric fluid is flooded into the spark gap.

When precision band sawing hard materials, a fine wire wound onto two reels is used, with the dielectric fluid flooding into the work. To prevent the wire from breaking and wearing it is continualy wound from one reel to the other.

6.2.2 Electrochemical machining

Electrochemical machining (ECM) is a process of metal removal in which electrolytic action is utilised to dissolve the work piece material. It is, in effect, the reverse of electroplating. The term 'ECM' is sometimes used for the electrolytic grinding process, which is a modification of electrochemical machining.

The basic elements of an electrochemical machine are shown in *Figure 6.2*. The work piece (which must be a conductor of electricity) is placed in a tank on the machine table and connected to the *positive* terminal of a d.c. supply. The tool electrode, which is shaped to form the required cavity in the work piece, is mounted in the tool holder and connected to the negative terminal of the supply. An electrolyte flows through the gap between the tool and work piece and is then pumped back to the working zone, either through the tool or externally, depending on the application.

The action of current flowing through the electrolyte is to dissolve the metal at the anode, i.e. the work piece. The electrical resistance is lowest (and hence the current highest) in the region where the tool and work are closest together. Since the metal is dissolved from the work most rapidly in this region, the form of the tool will be reproduced on the work.

There is no mechanical contact between work and tool and any tendency of the work metal to be plated on the tool (the cathode) is counteracted by the flow of electrolyte which removes the dissolved metal from the working zone. Hence there is neither tool wear nor plating of the work material on the tool, so that one tool can produce a very large number of components in its life.

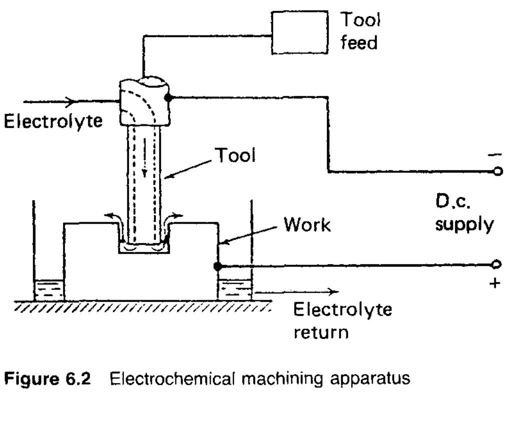


Figure 6.2 Electrochemical machining apparatus

6.2.2.1 Surface roughness

The fact that metal removal in ECM is not achieved by mechanical shearing or by melting and vaporisation of the metal means that no thermal damage occurs and no residual stresses are produced on the worked surfaces. The only heat generated is that due to electrical resistance, and the temperature cannot be allowed to rise above the boiling point of the electrolyte.

Surface finishes of 0.75–1 μm are readily obtained, while surface finishes better than 0.25 μm have also been achieved. A great advantage of ECM, which is contrary to conventional machining, is that, as the metal removal rate is increased, the surface finish and accuracy are improved.

6.2.2.2 Features

These can be listed as follows.

(1) Stress-free surface on work piece.
(2) No work-hardening of material.
(3) No thermal cracking of work surface.
(4) High surface finish.
(5) No softening of work surface.
(6) Freedom from burrs.
(7) High metal removal rates on hard materials.
(8) No electrode wear.

6.2.3 Electrolytic grinding

This is a modification of the ECM process. The tool electrode consists of a rotating abrasive wheel which can conduct electricity; this is usually a metal bonded diamond wheel. The electrolyte is fed between the wheel and work surface in the direction of movement of the wheel periphery, so that it is carried past the work surface by the wheel rotation. The abrasive particles help to maintain a constant gap between wheel and work. The current flows between work and wheel as shown in *Figure 6.3*.

In this process the predominance of the electrolytic action reduces wear to a negligible amount and makes it possible to grind hard materials rapidly. Furthermore, the wheel can be used for long periods without dressing.

In applying electrolytic grinding it is desirable to design the operation in such a way that the area of contact between the wheel and work is as large as possible. This gives the highest rate of removal for a given current density and allows full use to be made of the available current capacity. The most notable application of this method is in the grinding of carbide tool tips where surface finishes of the order of 0.1 μm have been obtained.

6.2.3.1 Applications

The main application of electrolytic grinding is in machining hard metals such as those used in high temperature service.

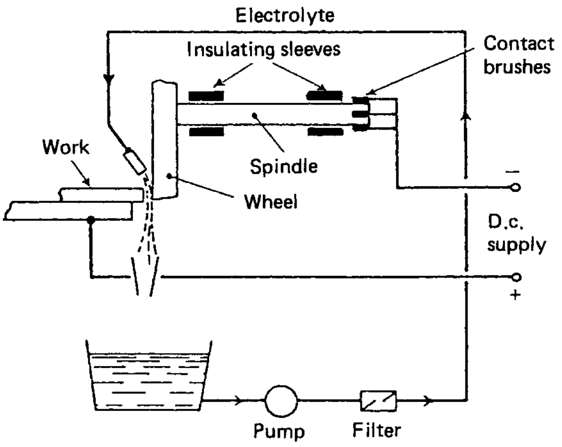


Figure 6.3 Electrolytic grinding apparatus

The removal rate is much higher than in conventional machining.

ECM has been applied successfully to the following machine operations:

(1) machining of through holes of any cross-section;
(2) machining blind holes with parallel sides;
(3) machining shaped cavities, such as forging dies;
(4) wire cutting of large slugs of metal; and
(5) machining complex external shapes, e.g. turbine blades.

6.2.4 Other electrochemical processes

A number of ECM processes find application in industry. These include:

(1) electrochemical turning (ECT);
(2) electrochemical honing (ECH);
(3) Shaped tube electrolytic machining (processes developed by General Electric Co., Aircraft Engine Group)—this process is used to drill small shaped or round holes (0.64–6.35 mm diameter), in electrically conducting materials it is usually difficult to machine alloys;
(4) Electro-stream (process developed by General Electric Co., Aircraft Engine Group)—this process is used to 'drill' extremely small holes (0.21–1.02 mm diameter);
(5) electrolytic polishing;
(6) electrolytic end milling; and
(7) electrolytic belt grinding.

6.3 Mechanical processes

6.3.1 Water-jet machining

Water-jet machining removes work-piece material and produces a narrow slit by the cutting action of a fine, high pressure, high velocity stream of water or water based fluid with additives. The water-jet nozzle can be designed to mix abrasive particles with the water and hence provide abrasive jet machining.

High pressure is needed to cut large thicknesses, while low pressure is used for deburring or finishing processes. The machined surface is generally unaffected by stress and heat.

6.3.1.1 Applications

Complex profiles may be produced using sophisticated nesting techniques when the nozzle is carried on a computerised numerical control cutting system. Computer control of the abrasive feed, water pressure and cutting speed ensures precise cutting.

Water jets are used for slitting and profiling of many non-metallic materials such as wood, paper, glass, marble, asbestos, plastics, leather, rubber, nylon and fibreglass. To cut large thicknesses of steel sheets, high pressure water jets are required.

6.3.2 Ultrasonic machining

Machining by the electrochemical or electrical discharge processes is confined strictly to metals, whereas ultrasonic machining may be applied to a much wider range of materials. Typical examples of the range of work covered by this process are:

(1) machining of ceramics and porcelain;
(2) cutting glass, including thread cutting;
(3) cutting precious and semiprecious stones;
(4) machining silicon, germanium and other semiconducting materials;
(5) machining ferrite;
(6) machining small punches and dies for presswork;
(7) machining accurately cavities of limited depth in cemented carbide; and
(8) drilling holes of varied internal configuration.

The development of this process has opened up several possibilities for the extended use of materials such as ceramics and semiconducting materials which would hitherto have been impractical. It must be recognised, however, that at this moment in time there are limitations to the process which have been not been overcome, namely:

(1) relatively small machining area;
(2) low cutting rates for hard metals relative to other work materials;
(3) relatively high power consumption;
(4) high rate of tool wear when cutting hard metals relative to other work materials; and
(5) relatively shallow depth of cut.

The cutting of brittle materials by the application of repeated blows of a concentrated nature has long been applied. This principle is applied in ultrasonic machining where the machining of a wide range of brittle materials can be carried out.

In ultrasonic machining the material is removed by the action of an abrasive which is driven into the work surface by a tool oscillating normal to the surface at high frequency. The tool which is machined to the shape of the desired hole or external profile is made from a tough metal, such as tool bits of alloy steel. A slurry of abrasive powder in liquid is fed to the tool which is oscillated at frequencies between 2000 and 3000 Hz with amplitudes in the region of 0.025–0.075 mm. The tool is pressed onto the work with a force of a few newtons, feeding down as the profile is cut on the work.

Excitation of the tool at these frequencies may be produced by an electromechanical transducer of the piezo-electric or magnetostrictive design.

6.3.2.1 Basic principle

The basic principle of ultrasonic machining is shown in *Figure 6.4* and, in order to increase the amplitude of vibration of the

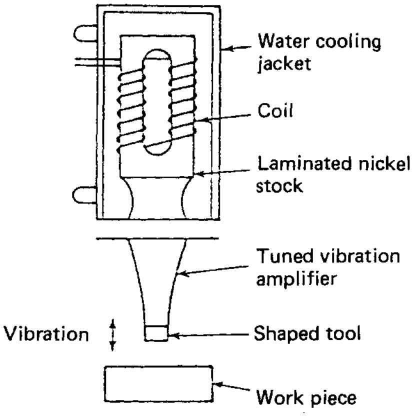


Figure 6.4 The principle of ultrasonic machining

transducer, a velocity transformer is fixed to the radiating face of the transducer. The resonant movement of the radiating face of the transducer being in the region of 0.012 mm, by attaching a velocity transformer the longitudinal amplitude can be increased to 0.025–0.075 mm. The transformer is made from materials having high fatigue strength and low energy loss and a material such as brass meets this specification. The profile of the transformer is tapered exponentially (*Figure 6.5*), the increase in the ratio of the amplitude being inversely porportional to the ratio of the areas of the two ends.

The toolholder (or tool cone) is fixed to the end of the transformer and to give minimum damping the vibrating components are clamped at the nodes and are acoustically matched. A transducer–transformer–tool system is thus produced which is a resonant mechanical transmission line that is two or more half-wavelengths long. As tool wear produces shortening of the combined length of tool cone and tool, it is usual practice to make these slightly longer than a half-wavelength and, thereby, increase their effective life.

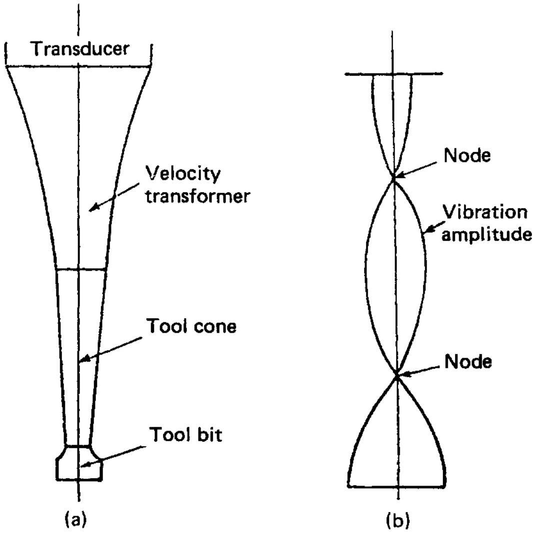


Figure 6.5 Profile of the transformer

When machining components by this method the tool is held against the work by a light static force of about 18 N and an abrasive slurry is applied to the work area either by hand or by a circulating pump. As the tool tip is oscillated by the transducer, abrasive particles are trapped between the tool and the work and a chipping action ensues which removes small pieces of material from the work piece. The vibration of the tool tip accelerates the abrasive at very high rates, thus imparting the necessary force to produce the cutting action. Cavitation which occurs in the gap assists removal of the chips and circulation of the abrasive.

The abrasive is suspended in a liquid which serves several functions; it provides an acoustic bond between vibrating tool and work and, therefore, produces an efficient means of transferring energy between tool and work. It also acts as a coolant at the tool face and assists in carrying the abrasive to the cutting area and to wash the spent abrasive and swarf away.

A number of different types of abrasive have been used for ultrasonic machining, but those most commonly used are aluminium oxide, silicon carbide, boron carbide and, to a lesser extent, diamond dust. The grit size varies between 200 and 2000 with the coarse grade being used for roughing and the finer grades (say 600–700) used for finishing. The very fine grades (1000–2000) only being used for the final pass on work of the highest accuracy.

The liquid used to carry the abrasive must have the following properties:

(1) its density must be approximately equal to that of the abrasive;
(2) it must have good wetting properties;
(3) in order to be able to carry the abrasive down between tool and work, it must have low viscosity; and
(4) for the efficient removal of heat from the cutting zone it must have high thermal conductivity and specific heat.

Water will satisfy most of these requirements with such additives as rust inhibitors and wetting agents.

6.3.2.2 Surface finish and accuracy

The abrasive grit size is a major factor influencing surface finish. When using, for example, boron carbide on glass, the roughness may vary from 5 μm for 100-mesh grit to 1–25 μm when a 600 mesh is used. If tungsten carbide is machined, however, the roughness values are approximately one-quarter of those listed above. It has been found that the roughness obtained on the face of the tool is generally less than that on the sides.

Tolerances as close as $\pm$0.013 mm may be readily attained.

6.3.2.3 Advantages

The advantages of ultrasonic machining are as follows.

(1) Accuracy and good surface finish are easily obtained.
(2) No heat is generated and, therefore, no changes occur in the physical structure of the material.
(3) Equipment is safe to handle and little skill is required in its operation.
(4) Cheap abrasives are used.
(5) The process is applicable to all brittle materials and is not limited by any other physical property.

6.4 Thermal processes

Thermal machining processes use thermal energy as the primary source of energy for metal removal. The energy is of such a magnitude as to evaporate the material being machined.

6.4.1 Electron-beam machining

Electron-beam machining (EBM) (*Figure 6.6*) is based on the principle that it is possible to accelerate electrons to very high velocities, i.e. upwards of 160 000 km/s. The electrons are then focused into a narrow beam by an electric field which can be bent by electrostatic and electromagnetic fields in a similar way to light rays being bent by glass lenses. The high-velocity stream of electrons are so focused as to impinge upon a very small spot on the work material, the kinetic energy of the electrons then being transformed into thermal energy which vaporises the material in the local area of the focused spot. In order to prevent the electrons from colliding with gas molecules in the atmosphere, which would scatter the electron beam, the operation is carried out under vacuum.

6.4.1.1 Advantages and limitations

This process provides what is perhaps the most precise cutting tool available. It can be used to cut any known material, metallic or non-metallic, which can exist in a high vacuum. As all cutting must be done under vacuum, the degree of vacuum restricts the component size, and the time must be allowed to evacuate the chamber. This method is very suitable for micromachining, as holes down to 0.05 mm diameter can be machined with no cutting-tool pressure or wear. Hole diameter/depth ratios of the order of 1:200 can be achieved and, due to the small beam diameter, close tolerances can be held (±0.012 to ±0.05 mm).

Distortion-free machining of thin or hollow parts is possible with no physical or metallurgical damage to the work material. Only relatively small cuts are economically possible, as the metal removal rate is only approximately 0.1 mg/s. The holes produced by this method have a slight crater where the beam enters the work and the hole is tapered by about 2° per side. The profile of the hole varies with its depth, due to the fact that the beam fans out above and below its focal point, thereby tending to produce an hourglass shape.

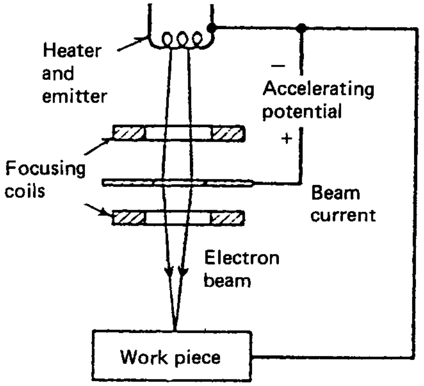


Figure 6.6 Electron-beam machining apparatus

High capital cost of the equipment plus the need for considerable operator skill and the fact that the process is applicable to thin material rather limits its range of application.

6.4.1.2 Applications

The process has been applied to such operations as drilling either round or profile shaped holes for metering flow on sleeve valves, and rocket fuel injectors of injection nozzles for diesel engines. Further applications have been concerned with producing wire-drawing dies, spinnerets for manufacturing synthetic fibres and drilling gas orifices for pressure-differential devices.

6.4.2 Laser-beam machining

Laser-beam machining comprises the conversion of electrical energy into light energy and then into thermal energy. The highly focused, high density energy melts and evaporates a portion of the work piece in a controlled manner (*Figure 6.7*).

The term 'laser' is an acronym for 'light amplification by stimulated emission of radiation'. Lasers which are capable of producing highly directive beams of visible, infra-red and ultraviolet radiation, can be classified as solid state, gas, or liquid lasers. The process is usually automated by using computerised numerical control systems to control the movement of the machine table under the laser beam.

6.4.2.1 Applications

The laser machining process does not require a vacuum, and most engineering materials can be successfully laser processed. These include mild, stainless and special steels, aluminium, brass, nimonics, titanium, acrylics, ABS and other plastics, rubber, leather, textiles, Kevlar, advanced composites, quartz glass and fibre glass.

The process is suitable for small-scale cutting operations such as slitting and drilling of holes which may be as small as 0.005 mm with depth/diameter ratios of 50:1.

The following is a guide to the thicknesses of materials which can be cut.

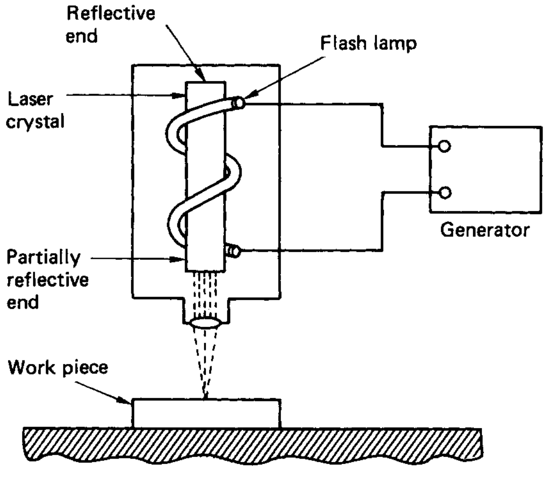


Figure 6.7 Laser-beam machining apparatus

Material	*Maximum cutting thickness* (mm)
Mild steel	6
Stainless steel	5
Aluminium	2
High temperature alloy	4
Brass	2
Acrylics	15

The main disadvantage in laser applications is the cost of the equipment.

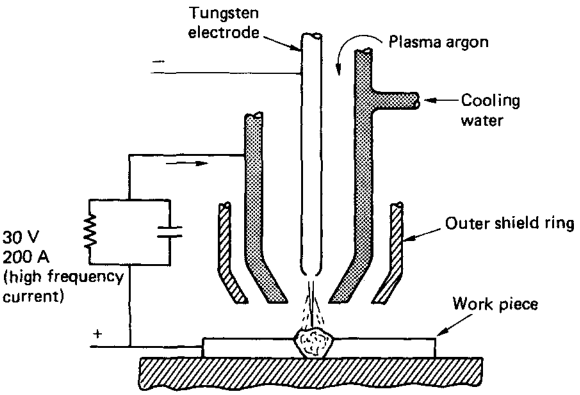


Figure 6.8 Plasma-arc machining apparatus

6.4.3 Plasma-arc machining

A 'plasma' is defined as a gas that has been heated to a sufficiently high temperature to become partially ionised and, therefore, electrically conducting. The temperature of the plasma may reach as much as 30 000°C. A torch is generally used to provide an electric arc between an electrode and the work piece. A typical plasma torch is shown in *Figure 6.8*.

6.4.3.1 Applications

The stream of ionised particles from the nozzle can be used to perform a variety of industrial jobs, including welding, cutting and machining.

Welding Plasma arc welding can be employed to join almost all metals. The process is generally applied manually and requires some degree of welding skill.

Cutting With appropriate equipment and techniques, the plasma arc can be employed to make cuts in electrically conducting metals. It can be used for cutting very hard materials and for machining of holes and profiles.

Machining Plasma arc methods are employed in special applications to replace conventional machining operations such as the removal of metal from the surface of a rotating cylinder to simulate the turning operation.

7 Electronic Manufacture

D Koshal

Contents

7.1 Introduction 7/3

7.2 Manual assembly of through-hole components 7/3
7.2.1 Low throughput 7/3
7.2.2 Medium throughput 7/3
7.2.3 Manual assembly with the aid of an assembly director 7/3
7.2.4 Automated assembly of through-hole components 7/4

7.3 Surface-mount technology 7/4
7.3.1 Benefits of surface-mount technology 7/4
7.3.2 Limitations of surface-mount technology 7/4

7.4 Surface-mount components 7/4
7.4.1 Active components 7/4
7.4.2 Passive components 7/5
7.4.3 Component packaging 7/6

7.5 Assembly of printed circuit boards 7/7
7.5.1 Single-sided PCB with SMCs 7/7
7.5.2 Double-sided PCB with SMCs on both sides 7/7
7.5.3 Mixed assembly, double-sided PCB with SMCs on one side only 7/8
7.5.4 Mixed assembly, double-sided PCB with SMCs on both sides 7/8

7.6 Surface-mount-component attachment methods prior to soldering 7/9
7.6.1 Adhesive dispensing 7/9
7.6.2 Curing the adhesive 7/9
7.6.3 Solder pastes 7/10
7.6.4 Flux 7/10

7.7 Soldering 7/10
7.7.1 Wave soldering 7/10
7.7.2 Reflow methods 7/11

7.8 Cleaning 7/12

7.9 Automatic testing equipment 7/12

7.10 Surface-mount-component placement machines 7/12
7.10.1 Selection criteria for placement machines 7/12
7.10.2 Types of machine 7/12

7.11 Printed-circuit-board layout 7/12

7.12 Possible defects during manufacture 7/13

7.13 Introducing surface-mount technology 7/13

7.1 Introduction

Electronic manufacture can be classified as the manufacture of printed circuit boards (PCBs) and electronic components, the placement of components onto PCBs, and the assembly of boards into a plastic/metal box or case. For several decades, through-hole components technology has been used to achieve great advances in the electronics industry, but now there is a growing trend whereby these components are being replaced by a method called 'surface mounting' in which the component leads or terminals are soldered to the top surface of the board.

This trend has accelerated recently due to the advantages offered by surface-mount technology (SMT) in terms of reduced cost, size and improved reliability. The new volume efficient packages for integrated circuits (ICs) and the use of passive chip components allow more dense circuit packing, permitting either an increase in functions for a given equipment size or a reduction in equipment size and lower system costs.

It has been found that approximately 40% of the cost of an electronic product is in the assembly of components onto a PCB. Therefore, an engineer needs to select the appropriate assembly method to minimise the cost of manufacture of a product. This chapter describes the various component assembly techniques that are currently used in industry.

7.2 Manual assembly of through-hole components

Figure 7.1 shows the stages involved in the manual assembly of PCBs with through-hole components. The types of assembly method currently used in industry depend on the quantity and the variety of boards required. There are three types of manual production which are commonly used. In the case of high throughput, automatic machines are available.

7.2.1 Low throughput

In this type of assembly an operator completes the entire board, including soldering. The board may consist of through-hole and surface-mount components. The reason for this type of production line is usually that either the board to be produced is very densely populated with components, or the board contains delicate or sensitive components.

7.2.2 Medium throughput

A cell is set up consisting of several operators and each person is trained to insert about 10 components manually into the PCB. When the board reaches the end of the line, the board is complete and, after visual inspection, is ready for wave soldering. To aid insertion of components a master board is usually made available to each operator.

7.2.3 Manual assembly with the aid of an assembly director

In the above methods the operators are required to memorise the component locations and it is inevitable that mistakes may be made. To reduce misinsertion of components, a machine called an 'assembly director', is employed (see *Figure 7.2*).

In the normal mode of operation of the assembly-director machine, the beam deflection system of the helium–neon laser displays on the surface of the board to be assembled one or more component locations. At the same time various component-dispensing machines can be driven to dispense the appropriate component. To advance to the next preprogrammed location a foot switch is incorporated, thus leaving the hands free for the insertion of components.

The method probably does not substantially increase the production rate, but it does reduce rework.

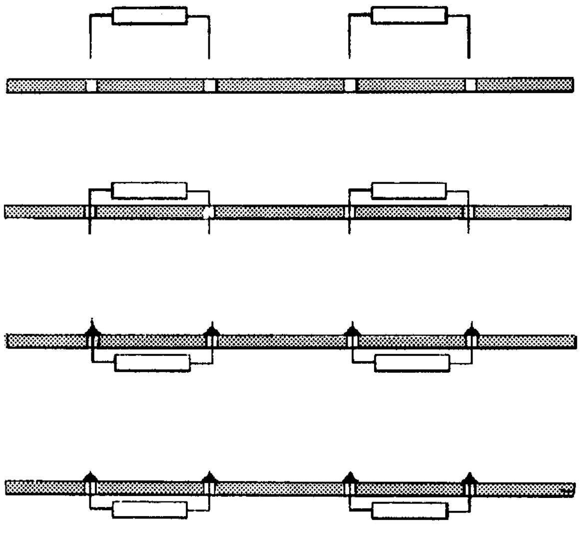


Figure 7.1 Stages in the assembly of through-hole components onto printed circuit boards

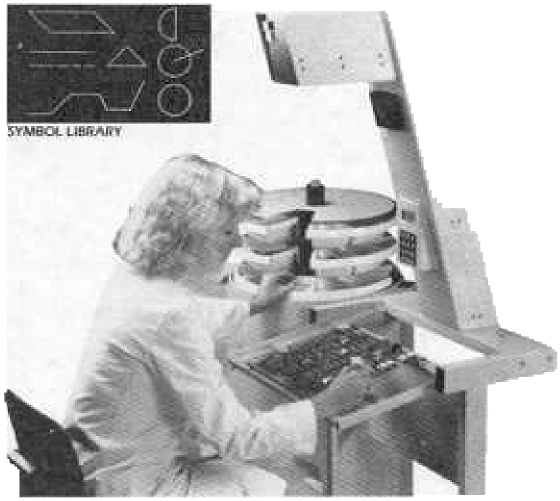


Figure 7.2 Assembly of components using an assembly director

7.2.4 Automated assembly of through-hole components

Automatic component insertion machines are available that are capable of inserting axial, dual in line (DIL), and radial components at a rate of about 1200 components per hour. Machines which are dedicated to inserting one or two types of the above components can operate at a much faster rate.

7.3 Surface-mount technology

Surface-mount technology (SMT) is defined as the placement of components on a printed circuit planar surface with component leads or terminals being soldered to that surface.

Surface-mount components (SMCs) are employed in various electronic products, including disc drives, personal computers, digital watches, communication equipment, test instrumentation, television, videos, and cameras.

7.3.1 Benefits of surface-mount technology

Circuit design

(1) An increase in board packaging density is achieved. The use of SMT components can result in a 70% improvement in density. Savings of 35–50% are routine.
(2) An improvement in frequency response is obtained. Conventional component leads act as antennas which generate electrical noise, thus the reduction in lead length afforded by surface mounting gives SMT a major edge in terms of external noise immunity.
(3) A reduction in size and weight for the same function.
(4) An improvement in shock and vibration characteristics due to lower mass.

Board manufacture and cost

(1) A reduction in the number of boards required.
(2) A reduction in material costs.
(3) A reduced number of holes is needed, i.e. the drilling cost is reduced.
(4) A lower number of conductor layers is required in the board structure.

Assembly of the PCB

(1) This is easy to automate.
(2) There is a reduction in inventory and manufacturing.
(3) Improved reliability is obtained.
(4) Increased manufacturing control is achieved.

Electronic components

(1) SMCs are potentially cheaper than the leaded equivalent.
(2) Surface-mounted devices are available in smaller overall dimensions.
(3) Surface-mounted devices are more reliable.
(4) The packaging of components makes storage easy.

7.3.2 Limitations of surface-mount technology

(1) There are some components which cost more in surface-mount packaging than in through-hole technology.
(2) Although the range of components available in SMT is wide, there are some families of components which are either difficult to find or not offered at all in SMT.
(3) The cost for test and repair is higher.
(4) There is a higher requirement for the assembly process.
(5) A larger number of process steps is required in assembly.

7.4 Surface-mount components

Surface-mount components are classified as follows:

Passive components
Resistors
Capacitors
Inductance
Connectors

Active components
Diodes
Transistors
Thyristors
Integrated circuits

7.4.1 Active components

Small outline integrated circuit (SOIC) packages (Figure 7.3(a)) These are currently available in 6, 8, 10, 14 and 16 pin versions with a body width of 4 mm and in 16, 20, 24 and 28 pin versions with a wider body of 7.6 mm. The leads of these devices are of the gull-wing style.

Small outline transistors (SOT) (Figure 7.3(b)) These are used for discrete semiconductors such as transistors and diodes. Three standard outlines are available. The SOT-23 and SOT-143 have leads on opposite sides, while the SOT-89 has all leads on the same side.

Leadless ceramic chip carriers (LCCC) (Figure 7.3(c)) These are the most commonly used type of package in surface-mounted applications. The package consists of a small piece of plastic or ceramic with etched circuit traces on both the top and the bottom with over-the-edge interconnections on all four sides. The wire bonded IC chip is on the top and the terminating pads are on the bottom where final attachment to the medium takes place. The LCCCs need no holes to be drilled through the PCB. They are commonly available in 18, 20, 28, 32, 44, 52, 68, 84, 100, 124 and 156 terminal versions.

Plastic leaded chip carriers (PLCC) (Figure 7.3(d)) These low-cost plastic leaded chip carriers have leads (gull-wing profile or J shaped) on all four sides of the package. They are square in outline and commonly have between 18 and 84 pins.

Quad flat packs (Figure 7.3(e)) These offer surface-mount packages for devices requiring 64 to 196 pins. The high pin count is achieved by reducing the pitch between the adjacent leads. The quad flat pack devices have one of three different lead pitches (0.65, 0.8 and 1.0 mm), depending on the number of pins required. The leads are formed to a gull-wing mode to bring the ends of the leads level with the bottom of the package.

Cylindrical diode packages (Figure 7.3(f)) These components are available in metal electrode face bonded packaging.

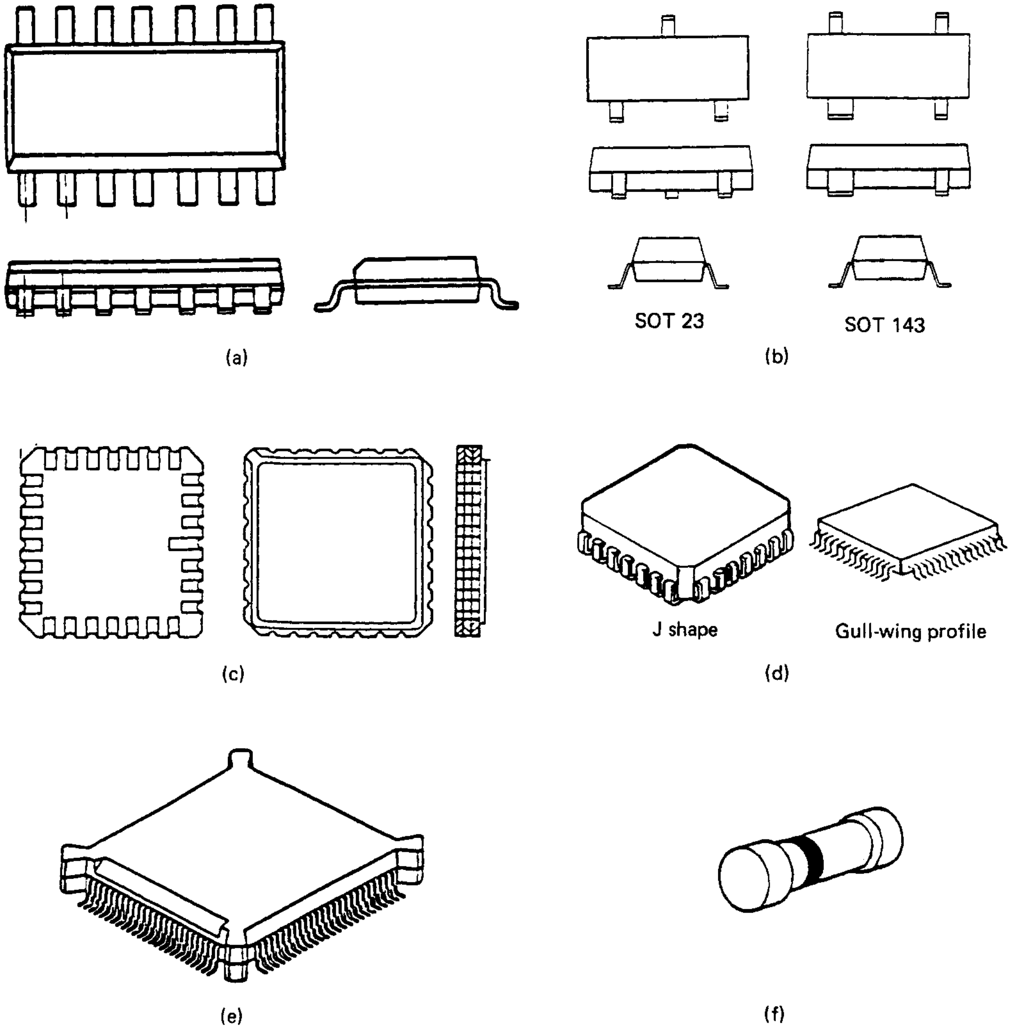


Figure 7.3 Active surface-mount components. (a) Small outline integrated circuit package. (b) Small outline transistors. (c) Leadless ceramic chip carriers. (d) Plastic leaded chip carriers. (e) Quad flat pack. (f) Metal electrode for bonded packages

Discrete diode components are frequently packaged in SOT-23 encapsulations, but one of the three contacts is then redundant. Several two-terminal packages have been developed especially for diodes, the two most popular (the SOD-80 and the MELF diode) both being cylindrical.

The SOD-80, sometimes referred to as the 'MiniMelf', is less costly than the SOD-23, requires less board space and weighs less. However, the power dissipation of the SOD-80 is limited to 250 mW. When more power dissipation is required, a larger cylindrical encapsulation is used.

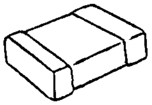
MELFs lead to placement problems because of the curve of the surface which will make contact with the board. The small area of contact means that the components can become detached relatively easily. The main advantages of MELFs is that they are less expensive than rectangular chips.

7.4.2 Passive components

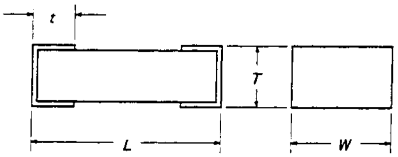
Chip resistors The most commonly used sizes are the 1206 with height of 0.6 mm and the 0805 with a height of 0.5 mm. The typical power dissipation is 0.125 W at 70°C. The mechanical data for these packages are shown in *Figure 7.4(b)*.

Chip capacitors Chip capacitors are described by the same size notation as used for chip resistors. The following types are available:

(1) ceramic chip capacitors,
(2) cylindrical ceramic capacitors,
(3) plastic capacitors,
(4) tantalum chip capacitors, and
(5) aluminium electrolytic capacitors.

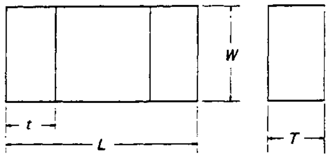


(a)



Size code	Dimensions (mm (in.)) L	W	T	t
0805	2.00 ± 0.20 (0.079 ± 0.008)	1.25 ± 0.20 (0.049 ± 0.008)	0.05 ± 0.20 (0.020 ± 0.008)	0.40 ± 0.25 (0.016 ± 0.010)
1206	3.20 ± 0.20 (0.126 ± 0.008)	1.58 ± 0.18 (0.062 ± 0.007)	0.56 ± 0.15 (0.022 ± 0.006)	0.50 ± 0.25 (0.020 ± 0.010)
1210	3.20 ± 0.20 (0.126 ± 0.008)	2.50 ± 0.20 (0.098 ± 0.008)	0.56 ± 0.15 (0.022 ± 0.006)	0.50 ± 0.25 (0.020 ± 0.010)

(b)



Size code	Dimensions (mm (in.)) L	W	T_{max}	t_{max}
0805	2.0 ± 0.2 (0.079 ± 0.008)	1.25 ± 0.2 (0.049 ± 0.008)	1.3 (0.051)	0.7 (0.028)
1206	3.2 ± 0.2 (0.126 ± 0.008)	1.6 ± 0.2 (0.063 ± 0.008)	1.5 (0.059)	0.7 (0.028)
1210	3.2 ± 0.2 (0.126 ± 0.008)	2.5 ± 0.2 (0.098 ± 0.008)	1.7 (0.067)	0.7 (0.028)
1812	4.5 ± 0.3 (0.177 ± 0.012)	3.2 ± 0.2 (0.126 ± 0.008)	1.7 (0.067)	0.75 (0.030)
1825	4.5 ± 0.3 (0.177 ± 0.012)	6.4 ± 0.4 (0.252 ± 0.016)	1.7 (0.067)	0.75 (0.030)

(c)

Figure 7.4 Passive surface-mount components. (a) Chip component. (b) Standard resistor sizes. (c) Standard capacitor sizes

The mechanical data for these packages are shown in *Figure 7.4(c)*.

7.4.3 Component packaging

In surface-mount assembly, the components must be packaged such that they can be handled by automatic equipment. The components must be presented to the assembly machine in a way that the machine tooling can locate and pick up the part. The commonly used types of packaging suitable for SMC placement machines are described below.

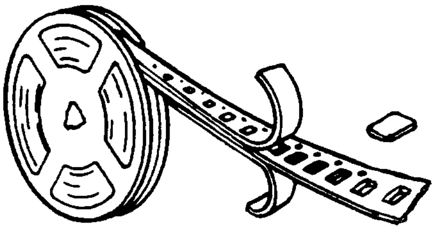
Sprocket driven tape (Figure 7.5(a)) 8, 12, 16, 24, 32, 44 and 56 mm sizes are available in tape which is either punched and formed cardboard or moulded plastic.

Bulk Bulk feeding is only possible with linear or vibratory feeders, specially tooled to handle a given part size. By means of a stop device fitted at a known position, the feeders can provide a source of components for the pick-up tool of a placement machine. Bulk purchasing is best suited to non-polarised components which are to be hand placed.

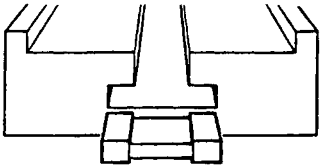
Sticks (Figure 7.5(b)) The components in flexible and rigid 'sticks' remain oriented and protected once loaded onto the machine. The machines are designed to select and remove devices from the sticks.

Matrix trays and waffle packs (Figure 7.5(c)) The internal waffle type nests (50 mm × 50 mm and larger) are made to conform to the part size. Each recess in the tray accommodates a component.

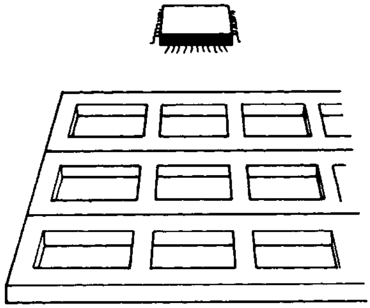
Figure 7.6 shows the body shapes and contact or lead arrangements, the circuit functions available, typical packages for each, and the general handling systems available.

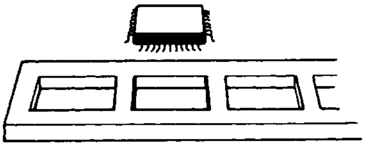


(a)



(b)





(c)

Figure 7.5 Component packaging. (a) Sprocket driven tape. (b) Sticks. (c) Waffle packs

Device function	Capacitor	Capacitor / Resistor	Resistor / Diode	Diode / Transistor	Transistor / Integrated circuit	Integrated circuit	Integrated circuit	Integrated circuit	Integrated circuit
Available package forms	Polarised chip	Chip	Cylindrical	SOT	Bare IC or flip-chip	SOIC	Plastic leaded chip carrier	Flat pack	Leadless chip carrier
Available handling systems									
Tape—8 mm and larger	●	●	●	●		●	●	●	●
Sticks – rigid and flexible			●			●			●
Waffle trays					●			●	
Bulk – vibratory feeder	●	●	●						

Figure 7.6 Typical component packages. SOT, small outline transistor; IC, integrated circuit; SOIC, small outline integrated circuit

7.5 Assembly of printed circuit boards

A typical assembly sequence of a printed circuit board (PCB) is as follows:

(1) insert pins/connectors,
(2) insert electronic components,
(3) solder PCB,
(4) clean PCB with solvent,
(5) inspection, and
(6) test.

Depending upon whether SMCs are used exclusively or in combination with through-hole components gives various configurations and assembly procedures (*Figure 7.7*).

The possible configurations are described below.

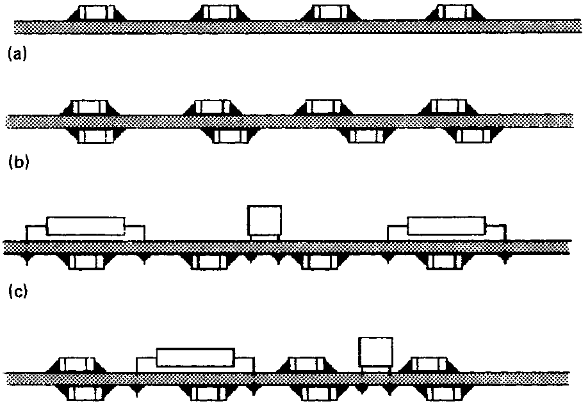


Figure 7.7 Various types of surface-mount assembly. (a) Single-sided printed circuit board (PCB) with surface-mount components (SMCs). (b) Double-sided PCB with SMCs on both sides. (c) Mixed assembly double-sided PCB with SMCs on one side only. (d) Mixed assembly double-sided PCB with SMCs on both sides

7.5.1 Single-sided PCB with SMCs

Figure 7.8 shows the steps involved in the assembly of a PCB by wave soldering or by a reflow technique. Depending on the production volume of the boards, the adhesive or solder paste may be either screen printed or dispensed by the machine.

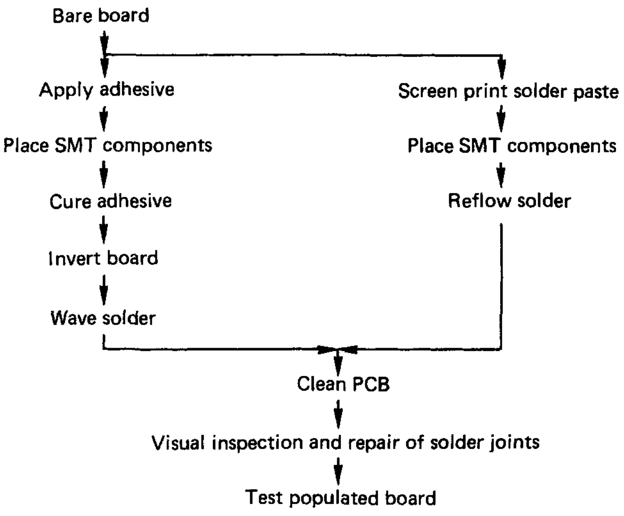
In the case of a sequential placement machine, the board is located on the table and the adhesive dispensing head places small drops of adhesive in the centre of the pads on which components are to be located. The pick-up head then places the SMCs on top of the adhesive. The adhesive is tacky enough to handle the populated board, but it should not be subjected to excessive vibration.

After visual inspection of the board it is placed in an oven to cure the adhesive. The PCB can now be soldered using a wave soldering machine. If solder paste is used it may be screen printed onto the board. After placing the SMCs the board is heated in an infra-red heater to melt and flow the solder to form a joint.

7.5.2 Double-sided PCB with SMCs on both sides

Considerable savings can be gained by mounting SMCs on both sides of the board using through-plated vias to make the interconnections between the opposite sides.

The steps involved are to screen print solder paste onto one side of the board, mount SMCs, and reflow the solder. Then, on the other side of the board apply the adhesive, place the SMCs, cure the adhesive, invert the board, and flux and wave solder (*Figure 7.9*).



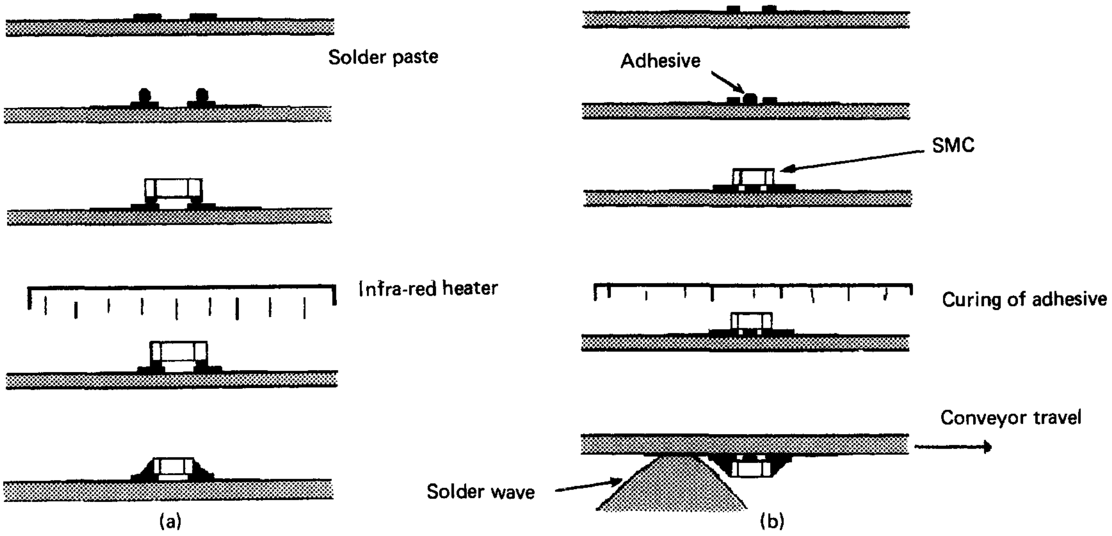


Figure 7.8 *Stages in the assembly of a single-sided printed circuit board (PCB) with surface-mount components (SMCs) using (a) solder paste and (b) adhesive attachment*

7.5.3 Mixed assembly, double-sided PCB with SMCs on one side only

In this case it is a common practice to place the SMCs before the through-hole components. The adhesive is dispensed, SMCs are placed, and the adhesive is cured. The board is turned over, through-hole components are inserted manually or by machine, and the board is fluxed and soldered (*Figure 7.10*). In this case the SMCs will be fully submerged in the solder paste. Therefore temperature-sensitive components will need to be assembled manually after soldering.

However, it is possible to assemble the through-hole components prior to the SMCs, but the above-described technique has the advantage that the adhesive can be screen printed, which is a much faster process than dispensing adhesive by means of a sequential placement machine.

7.5.4 Mixed assembly, double-sided PCB with SMCs on both sides

Figure 7.11 shows the steps involved in the assembly of SMCs and through-hole components on both sides of the board. Solder paste is screen printed onto the top surface of the PCB and the SMCs are placed using a sequential placement machine. Solder paste is then reflowed in an infra-red heater. The through-hole components can then be inserted manually or by means of a machine. The board is turned over, the adhesive applied, the SMCs placed, and the adhesive cured. The board is then turned over again and wave soldered.

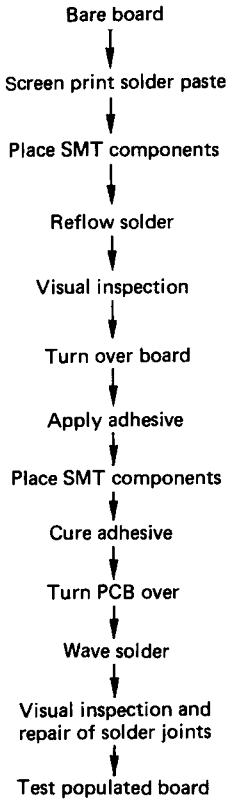


Figure 7.9 The steps involved in the assembly of a double-sided printed circuit board (PCB) with surface-mount components (SMCs) on both sides

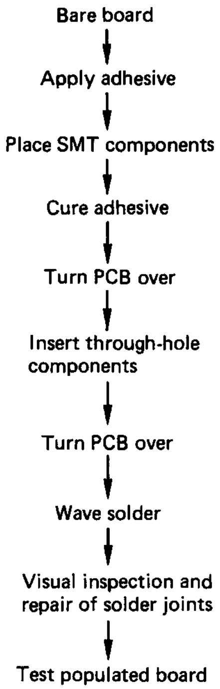


Figure 7.10 The steps involved in the mixed assembly of a double-sided printed circuit board (PCB) with surface-mount components (SMCs) on one side only

7.6 Surface-mount-component attachment methods prior to soldering

For flow soldering, SMCs must be held in place by an adhesive until they are soldered. The adhesive is applied to the component sites prior to SMC placement, and is cured immediately after placement. Manufacturing speed is dependent on the adhesive being quickly cured. SMT flow soldering subjects the SMCs to one or more passes through molten solder while hanging from the bottom of the board (*Figure 7.12*). Therefore, it is important that the cured adhesive has sufficient holding power for the task in hand, and that it retains the holding power at the elevated temperatures present during the soldering process.

Adhesives are generally left in place after the soldering process. Thus, long-term reliability depends on the adhesive being fully cured.

7.6.1 Adhesive dispensing

There are three methods by which a small amount of adhesive can be applied at each site where a surface-mounting device is to be placed.

Screen printing This method employs a polyester or stainless-steel screen consisting of a rigid frame on which is a stretched mesh. The screen is held at a distance from the substrate to which the adhesive or cream is to be applied and forced across the screen. Screen printing has the advantage of speed, but it cannot be used after conventional components have been inserted through the board.

Syringe A preset quantity of adhesive can be applied via a dispenser. The quantity of adhesive can be controlled by the adjustment of air-supply pressure, pressure-pulse duration, viscosity of adhesive, and the internal diameter of the nozzle. Dispensing the adhesive from a syringe is feasible for small boards with an assembly rate of less than abut 5000 components per hour. The syringe can either form part of the programmed pick-and-place machine or can be hand held for prototype manual assembly or repair work.

Pin transfer A pin is raised from below the surface of the adhesive and the small amount of adhesive at the tip of the pin is transferred to the substrate. With this method, either single pins can be used sequentially around the board or, for larger boards, an array of pins can be used for simultaneous application.

7.6.2 Curing the adhesive

Spots of adhesive are applied to the board followed by the placement of components upon it. At this stage the adhesion must be adequate to hold the component in place while the board is being handled. The adhesive is then hardened by curing.

To cure the adhesive the following methods are commonly used.

(1) Heating in an oven—a typical curing time for epoxy formulations might be about 6 min at 150°C.

Bare board
↓
Screen print solder paste
↓
Place SMT components
↓
Reflow solder
↓
Insert through-hole components
↓
Turn PCB over
↓
Apply adhesive
↓
Place SMT components
↓
Cure adhesive
↓
Turn PCB over
↓
Wave solder
↓
Visual inspection and repair of solder joints
↓
Test populated board

Figure 7.11 The steps involved in the mixed assembly of a double-sided printed circuit board (PCB) with surface-mount components (SMCs) on both sides

(2) Ultraviolet radiation—in an ultraviolet conveyor system the typical curing time may be about 3 min at 150°C for acrylic formulations.

7.6.3 Solder pastes

Solder paste or solder cream consists of a powdered solder alloy bound in a flux medium. The paste holds the surface-mount components in place until the solder is reflowed, flux cleans the surface and promotes wetting, and the solder forms the final joint.

It can be applied in one of several ways. The most common method for volume production is screen printing, but other methods such as stencilling, dispensing from a syringe, roller coating and pin transfer are used.

7.6.4 Flux

Flux is used to prevent the formation of oxides or other compounds of the metal with components of air. Fluxes for soldering electronic assemblies have three main functions:

(1) to make the surfaces to be joined wettable and active;
(2) to protect those surfaces from contact with air, in order to prevent oxidation of the molten solder; and
(3) to promote wetting of the surfaces to be joined by modifying the surface forces which control the wetting process.

Flux is applied to the surfaces to be soldered either by conventional wave soldering methods such as foaming and waving or as a flux paste.

Fluxes currently used for soldering fall into the following three main classes:

(1) rosin-based fluxes,
(2) synthetic fluxes, and
(3) organic fluxes.

7.7 Soldering

7.7.1 Wave soldering

Wave soldering can be done by the single-wave, dual-wave or the conventional drag-soldering approach. In dual-wave systems, the first wave is a turbulent wave which more evenly distributes the solder around the component. The second wave is a smoother wave which finishes the formulation of fillet and removes the excessive solder.

Wave soldering machines comprise a conveyor which transports the populated board either continuously or in a stepwise fashion from a loading position to, in turn, a fluxing station, a preheat stage, the solder wave and a cooling station, before removal from the conveyor at the unloading point. A gradual preheating profile is required to minimise the damaging effect of thermal shock (*Figure 7.12*).

Soldering of surface-mounting assemblies is carried out at a temperature in the range 235–260°C with a contact time of 1–4 s. The SMCs which move through the molten solder must be able to withstand this treatment, and remain unaffected either by the high temperature or the temperature gradients involved.

Advantages

(1) Suitable for soldering both through-hole and surface-mount components.
(2) Low technical risk when used on passive components.
(3) Especially suitable with a high percentage of conventional components.

Disadvantages

(1) SMCs are immersed in solder and hence exposed to large thermal gradients. They are not therefore suitable for temperature-sensitive surface-mount ICs.
(2) Difficulty in wetting all joints when components are closely spaced.
(3) Adhesive dispensing is required prior to placement.
(4) Dross is produced.

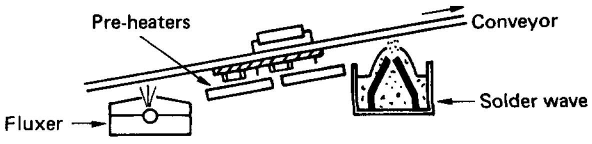


Figure 7.12 Wave soldering system

7.7.2 Reflow methods

Advantages

(1) Suitable for soldering a wide variety of surface components.
(3) Thermal gradients are easily controlled.
(3) The process is insensitive to component geometries.

Disadvantages

(1) Requires heavy investment in new equipment and process.
(2) There is a greater difficulty in adding conventional components.

7.7.2.1 Infra-red soldering

The process consists of a tunnel constructed from emitter lamps and refractory material. Both the mesh- and meshless-type conveyor holds the board and travels through the tunnel. The first segment of the system incorporates a short preheating zone which draws off flux volatiles in the solder paste. The next segment consists of a vent to extract these volatiles. The main area comprises the heaters with various controlled zones of top and bottom emitters. After a proper soldering profile has been established, the system is very repetitive (*Figure 7.13*).

Advantages

(1) Progressive and programmable heating rates.
(2) Adjustable solder temperature.
(3) Limited tombstoning.
(4) Low-cost operation.

Disadvantages

(1) For SMCs only.
(2) Solder paste dispensing required.
(3) Uneven temperature across the board.
(4) Temperature-sensitive components may require manual soldering.
(5) Some components such as J-bend and leadless chip carriers are not transparent to infra-red radiation.

7.7.2.2 Vapour phase reflow method

The vapour phase reflow method (*Figure 7.14*) involves passing the assembly to be soldered through a heated vapour. When a vapour comes into contact with the PCB below its condensation point, the vapour releases its latent heat of

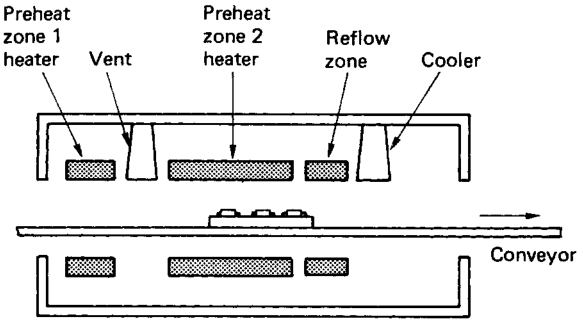


Figure 7.13 Infra-red reflow soldering system

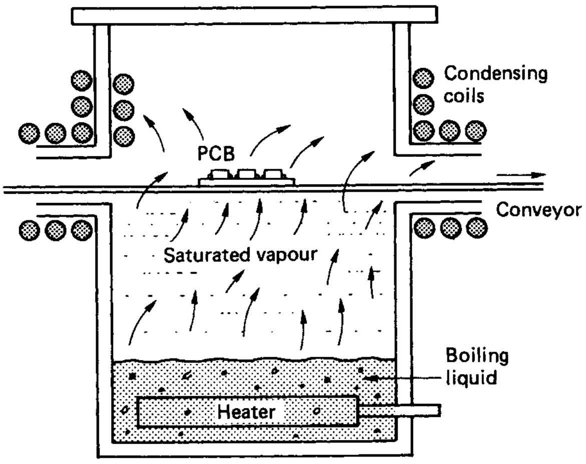


Figure 7.14 Vapour phase reflow soldering system. PCB, printed circuit board

vaporisation and condenses to form discrete droplets on the surface. As the vapour condenses on the cooler assembly it raises the temperature of the assembly to the point where the solder paste melts into solder. This process typically takes 15–30 s.

The fluid used in this process is usually one of a series of vaporised fluorocarbons with boiling points ranging from 174°C to 253°C.

Advantages

(1) Can be used for high density boards.
(2) The entire assembly is brought up to a well-controlled, uniform temperature where all solder connections are reflowed almost simultaneously.
(3) There is no large thermal gradient during heating.
(4) Components on both sides are soldered at the same time.
(5) Less flux is required because the soldering process takes place in an inert atmosphere.
(6) Components on both sides of the board can be soldered in one operation.

Disadvantages

(1) For SMCs only.
(2) Solder-paste dispensing required.
(3) The use of expensive chemical fluids is required.
(4) The high heating rates may damage some components.
(5) Mild acid formation occurs due to the decomposition of the working fluid.
(6) Tombstoning occurs.

7.7.2.3 Laser reflow soldering

With laser reflow soldering, each joint is made sequentially instead of simultaneously. Carbon dioxide lasers are used to heat the solder joints. Because the energy beam of the laser is highly focused, it is well suited to the soldering of temperature-sensitive components as well as narrow-spaced leads.

Advantages

(1) Only the lead solder paste is heated.
(2) Can be used for temperature-sensitive components.
(3) Produces high quality solder joints.

Disadvantages

(1) Low throughput because each joint must be reflowed individually.
(2) Very high equipment cost.

7.7.2.4 *Conductive belt reflow soldering*

In this process a flat heat-conducting belt is drawn over a controlled heated surface. The PCBs must have a very flat bottom surface and they are placed on the belt and moved through the heated area.

Advantages

(1) Low equipment cost.
(2) Simple operation.
(3) Moderate control over temperature profile.

Disadvantages

(1) Incompatible with organic PCBs.
(2) Small maximum board size.

7.8 Cleaning

The main function of cleaning an electronic assembly after soldering is to remove flux residues. Pure non-active resin flux should, by preference, be left unwashed on the board, but more usually an activated flux is required for good soldering and consequently washing is normally required. If water soluble flux is used, washing is mandatory. If washed with water, the subsequent drying process is expensive. Resin fluxes are washed away with solvents that are capable of removing both ionic and non-ionic contaminants from the assembly.

The type of flux used in wave soldering or the type of flux contained in the solder paste is very important in deciding whether or not to clean the assembly.

7.9 Automatic testing equipment

With the use of SMT there are fewer joints or solder joints to probe. The approach adopted has been to divide the board into 'modules'. These modules are tested on their testers in a quasi-functional test. The module is made small enough so that the fault can be narrowed down to a specific component in the module.

It is important to test the bare board as a large number of faults are due to faults in the board itself. It is less expensive to detect and correct such a fault in an unpopulated board than in an assembled one.

7.10 Surface-mount component placement machines

7.10.1 Selection criteria for placement machines

In selecting a placement machine, the following criteria must be considered.

(1) the PCB size to be assembled;
(2) the number of different PCB designs and batch sizes;
(3) the number of component types;
(4) the ease of programming the machine;
(5) the supply packing of components;
(6) the layout restrictions on the PCB;
(7) the testing of components for identity, polarity and faults;
(8) assembly performance speed;
(9) placement accuracy and reliability;
(10) assembly costs;
(11) adhesive dispensing; and
(12) flexibility.

7.10.2 Types of machine

There are four different placement methods used for SMCs.

In-line placement The PCB moves on a conveyor under several placement stations. Each module is dedicated to the placement of one type of package. Boards are loaded sequentially as they progress down the line. As a separate machine is required for each component, this approach is economical only for boards with a small number of components. Reconfiguration of the machine is difficult. In-line placement is best suited to the manufacture of a single product in very high volumes.

Sequential placement One or two software controlled pick-and-place heads pick up a SMC from a feeder or a feeding station and place it on the PCB. On some machines, the head moves in both x and y directions, while others employ a movable table to place the PCB under the fixed head. The latter technique is currently the most commonly used placement method. The accuracy of placement tends to be ± 0.1 to ± 0.2 mm for the x and y directions and $\pm 1°$ to $\pm 2°$ for rotation.

Simultaneous placement Multiple heads transfer an array of devices onto the board at one time. A board, or part of a board, is populated in a single operation. In this approach strict rules must be observed in the design of the board. The components must be laid out on a grid format which corresponds to the available feeder locations on the machine. The maximum number of components cannot exceed the number of feeder locations. Reconfiguration of the machine is usually very time consuming.

Sequential/simultaneous placement These machines incorporate an x–y table and multiple heads in succession. The heads are placed on a horizontally rotating turret. Each head picks up components at the feeding station and places them on the PCB.

7.11 Printed-circuit-board layout

The following factors affect the PCB layout.

Orientation The types of component used determine the choice of soldering methods available and thus influence the PCB layout. All components should be laid out parallel to the grid lines. This can result in an increased placement rate because of the reduction in the need for rotation of the pick up tool. The physical orientation of and the spacings between the components are primarily dictated by the soldering method to be used. Reflow soldering has no directionality and so components can be placed at any orientation to each other with a proximity restricted only by the footprint pad size requirements. Wave soldering, however, is directional and access to

the pads by the solder wave is required. Thus both the relative orientation of components and their spacing are of prime importance to the reliability of the soldering process.

Component density The component placement equipment that is to be used can impose restrictions on the component density, since some machines require vacant space for placement tools and space for rotating components. Other factors include:

(1) the component placement accuracy of the machine; and
(2) the clearance required for cooling the components, visual inspection and the use of test and rework equipment.

Soldering method The soldering method employed is also a major consideration in the design of the PCB artwork. When adhesive is used to attach a component, a dummy land is placed under the component between the pads on which the component is to sit. This land reduces the effective component standoff height and controls the spread of adhesive.

Test points Testing requirements may require special considerations such as test pads in the circuit and spacing between components that is sufficient to allow the use of test probes. It is general practice to add a solder-coated test pad to the conductor adjacent to the component.

7.12 Possible defects during manufacture

Skip This arises when there is no continuous solder between the substrate and the component and is due to a shadowing effect of the solder pads by a component body as the board passes through the solder wave.

Blow holes This results from the volatilisation and evolution of moisture absorbed in the epoxy of the board laminate. These faults are manifested as visible blow holes in the surface of the solder fillets, internal hidden voids within the fillets and considerable, or even total, blow-out of the solder from the holes.

Solder bridges These are unwanted solder paths between two or more conductors that should not be short circuited. Bridging is caused by excessive solder deposit, poor component placement, inadequate reflow time, or poor screening alignment. The occurrence of bridging can be reduced by changing the mechanism associated with the backwash region of the ware or by a change in the conveyor speed.

Poor solder Poor solder joints arise due to the incorrect setting of the wave height for the land. The solder does not come through the hole in a through-hole board.

Tombstone The 'tombstone', 'Manhattan', 'Stonehenge' or 'drawbridge' effect arises when a chip component is reflow soldered using solder paste, and one end lifts off the substrate during the reflow process. It is caused by using an excessive heating rate, uneven solder deposit, or poor component–pad solderability.

Misalignment of components In this defect the component is displaced sideways or rotationally with respect to its footprint on the PCB. Misalignment can be caused by a placement machine, or a subsequent movement prior to, or during, soldering.

7.13 Introducing surface-mount technology

The introduction of SMT must be part of a valid business strategy aimed at achieving a competitive edge. The impact of SMT is felt in every department. Therefore, teamwork and cooperation is essential from all concerned. The design and development requires computer-aided design systems with new capabilities. New design guidelines for SMCs need to be developed. Material management must develop new procurement methods and meet new goods inspection, storage and stock-control requirements. The manufacturing department must select, purchase and install new production lines for prototype and volume production, gaining new expertise in placement, reflow soldering, and cleaning. The manufacturing and test department must learn quickly and aim for zero defects. There is need to work together with people who will be required to control the successful introduction of the new technology.

The following three approaches have been used by many companies in introducing SMT products.

(1) *Minimum-risk approach*—this involves redesigning an existing board. The board often results in using both through-hole and surface-mount components.
(2) *High-risk approach*—a new product is introduced that is entirely surface mounted. This method is likely to give rise to a successful economic marketing edge.
(3) *Contracting out*—contracting out the surface-mount design to specialised companies is often adopted by smaller companies to evaluate the product in the market before investing in the SMT equipment. This option results in long-term benefits rather than short-term ones.

8 Metal Finishing Processes

T Buttery

Contents

8.1 Introduction 8/3

8.2 Abrasives 8/4
8.3 Grinding wheels and grinding wheel selection 8/4
8.3.1 Grit size 8/5
8.3.2 Wheel grade 8/6
8.3.3 Wheel structure 8/6
8.3.4 Bond 8/6

8.4 Mounting the grinding wheel 8/6

8.5 Balancing and dressing 8/7

8.6 Grinding mechanics 8/9

8.7 Wheel wear 8/11
8.7.1 Attritious wear 8/11
8.7.2 Fracture 8/11
8.7.3 Bond post-rupture 8/11

8.8 Grinding ratio 8/11

8.9 Grinding forces 8/11

8.10 Coolant 8/12

8.11 Grinding processes 8/12
8.11.1 Surface grinding 8/12
8.11.2 Work clamping 8/13
8.11.3 Cylindrical grinding 8/13
8.11.4 Centreless grinding 8/15
8.11.5 Tool and cutter grinding 8/17
8.11.6 Cutting off 8/17

8.12 Newer abrasives and grinding techniques 8/17

8.13 Honing 8/18
8.13.1 Basic principles 8/18

8.14 Honing practice 8/19
8.14.1 Abrasives 8/19
8.14.2 Tooling 8/19
8.14.3 Honing fixtures 8/20
8.14.4 Sizing systems 8/20
8.14.5 Machine configuration 8/20
8.14.6 Effects of honing 8/20

8.15 Superfinishing 8/23
8.15.1 Machining set-up and machining mechanisms 8/23

8.16 Coated abrasives 8/25
8.16.1 Cutting mechanism 8/26

8.17 Machining with coated abrasives 8/27
8.17.1 Abrasive belts 8/27
8.17.2 Abrasive belt selection 8/27
8.17.3 Contact wheels 8/28
8.17.4 Factors controlling abrasive belt performance 8/28
8.17.5 Abrasive belt machines 8/29

8.18 Abrasive discs 8/30

8.19 Lapping 8/30
8.19.1 Abrasives 8/30
8.19.2 Lapping processes 8/31

8.20 Polishing 8/33
8.20.1 Micropolishing 8/33
8.20.2 Barrel polishing/finishing 8/33
8.20.3 Barrelling medium 8/33
8.20.4 Barrelling compounds 8/33
8.20.5 Equipment 8/33

8.21 Blasting processes 8/33
8.21.1 Blasting abrasives 8/34

8.1 Introduction

In many applications, the functional behaviour of a component will depend critically on its surface and the regions immediately below. Factors contributing to surface performance are geometric shape, surface roughness R_a, materials properties, surface treatments and residual stresses. The total quality of a surface is, therefore, a function of the machining process by which it was produced and any surface treatment which may have been applied before or after machining to improve specific properties such as fatigue, wear or corrosion resistance.

Invariably, the approach to obtaining a 'perfect surface' involves a high cost penalty due mainly to the low rates of stock removal and high machining costs associated with abrasive machining processes (*Figure 8.1*). Continual efforts are therefore being made to improve metal removal rates and reduce floor-to-floor times for processes such as grinding, linishing and honing. Improved machining techniques, enhanced process control (by microprocessor or computer), in-process gauging and the application of automation or robotics increase the effectiveness of abrasive machining methods. Even though grinding gives good geometric shape, excellent dimensional tolerances and low values of surface roughness, it may not be ideal for some applications. For example, a ground surface has an approximately Gaussian height distribution (*Figure 8.2*), which in bearing applications will wear rapidly during the running-in period. Processes such as plateau honing are therefore applied after grinding to form a 'composite surface' having a topography similar to that of a component which has already been run in. In addition to the machining processes used to manufacture a component surface, treatments can also have major effects on surface quality and performance. A wide variety of processes is available to the engineer, ranging from the purely decorative to methods

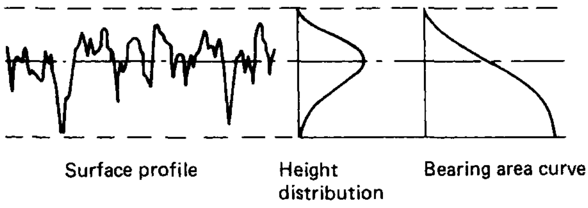


Figure 8.2 Surface profile of a ground surface together with its height distribution and bearing area curve

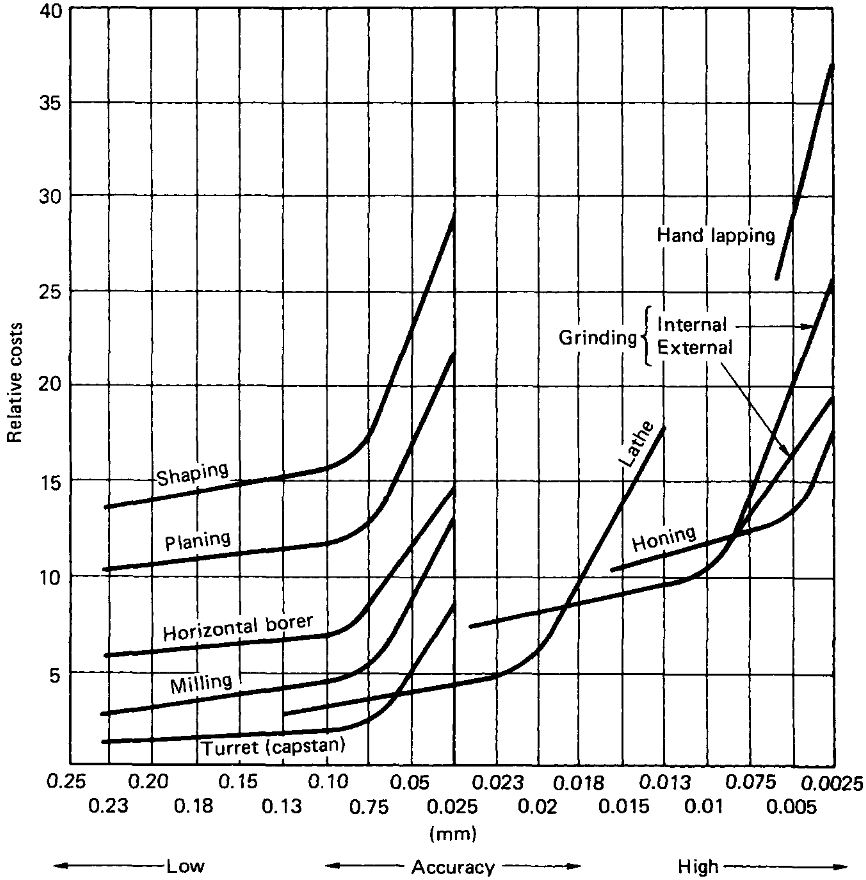


Figure 8.1 Relationship between the relative cost and accuracy for a number of machining processes. (Courtesy of British Standards Institution)

of increasing hardness and improving wear, fatigue, corrosion and scaling resistance. These methods vary in both complexity and cost.

8.2 Abrasives

In order for a material to act as an abrasive it must be significantly harder than the work piece. Richardson[1] has suggested a similarity with the Moh hardness scale, which places materials in 10 groups (0, soft, talc; to 10, hard, diamond) a material in a higher numbered group being able to scratch all materials in lower groups. Richardson concluded that an abrasive must be at least 10% harder than the work piece if it is to be effective. Furthermore, for an abrasive to act as a cutting tool it must have a suitable geometric shape, e.g. a sharp grit will cut more effectively than a blunt one, and as a hardened steel can act as a cutting tool or as a part of a bearing pair.

Many materials can be used as abrasives, but the most well known are silicon carbide (corundum), aluminium oxide (alumina), chromium oxide, iron oxide, magnesium oxide (magnesia), tungsten carbide, diamond (natural and synthetic) and cubic boron nitride. Abrasives are marketed in a variety of forms: solid abrasives such as wheels and sticks, coated abrasives such as belts and discs and loose abrasive in a range of particle sizes. For a particular application, the appropriate form must be chosen.

Although abrasives were originally naturally occurring, most of the abrasives in current use are man made. Silicon carbide and aluminium oxide were developed towards the end of the last century; more recently synthetic diamond (1950s) and cubic boron nitride (1960s) have become available. *Table 8.1* gives the hardness of the most widely used abrasive materials together with their chemical compositions.

There are two types of abrasive operation: two-body abrasion and three-body abrasion.

Two-body abrasion This type of abrasion involves interaction between the abrasive and the work piece. The abrasive is in solid form (grinding wheel, disc, stick or belt) and the work piece is forced against the abrasive making the second constituent of the abrasive pair (*Figure 8.3(a)*). Material is removed from the work piece as fine particles. Shot and vapour blasting are also examples of two-body abrasion; the abrasive pair consisting of the work piece and the abrasive particles propelled at its surface (*Figure 8.3(b)*).

Table 8.1 The hardness, chemical composition and applications of the most widely used abrasive materials

Material	*Chemical composition*	*Knoop hardness* ($kg\ mm^{-2}$)	*Applications*
Diamond	C	7000	Grinding tungsten carbide, glass and some die steels
Cubic boron nitride	BN (cubic)	4500	Grinding tool steel
Silicon carbide	SiC	2400	Used on cast iron, copper alloys, aluminium alloys and some steels
Aluminium oxide	Al_2O_3	2100	Softer but tougher than silicon carbide; used to grind steels
Sand/quartz	SiO_2	850	Abrasive belts and papers, sand blasting

Three-body abrasion Loose abrasive in a carrier, such as oil or paraffin, is sandwiched between two surfaces (*Figure 8.3(c)*) and, as the surfaces move, the abrasive rolls between them occasionally removing material. Cutting rates are therefore low and the method is used mainly to lap in mating surfaces or to produce very high quality finishes. When geometric shape is not important, a cloth may be used to carry the abrasive slurry; this method is widely used in the preparation of microspecimens.

8.3 Grinding wheels and grinding wheel selection

A conventional grinding wheel is effectively a composite material consisting of abrasive, bond and void; each of these components has a major influence on grinding-wheel perfor-

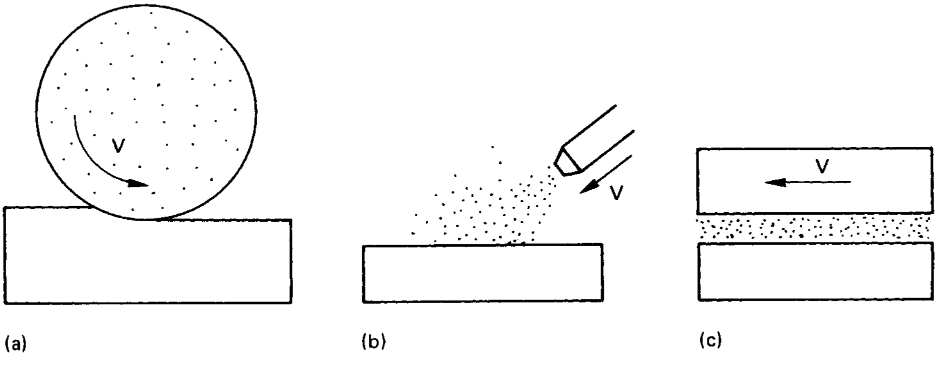


Figure 8.3 (a) Surface grinding. (b) Shot or vapour blasting. (c) Lapping

mance. The density of a grinding wheel (ρ_{wheel}) can be related to that of its constituent parts using the law of mixtures:

$$\rho_{wheel} = V_a\rho_a + V_b\rho_b + V_v\rho_v \qquad (8.1)$$

Since $\rho_v = 0$, equation (8.1) can be rewritten as

$$\rho_{wheel} = V_a\rho_a + V_b\rho_b \qquad (8.2)$$

where V_a, V_b and V_v are the volume fractions of the abrasive, bond and void, respectively, and ρ_a, ρ_b and ρ_v are the densities of the abrasive, bond and void, respectively.

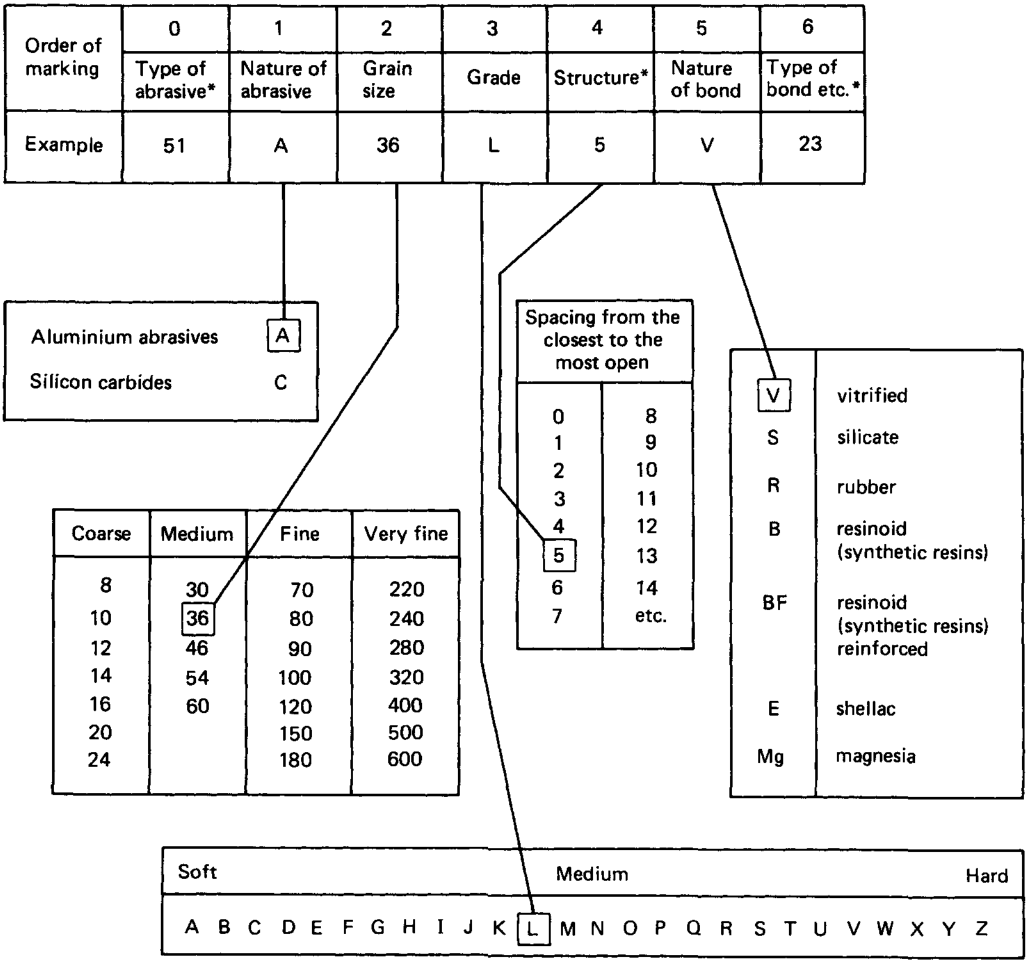
All the major properties and the ultimate performance of a grinding wheel are related to the proportions of the constituent parts and their individual characteristics.

The convention for the specification and identification of grinding wheels is set out in BS 4481: 1969: Part 1. Grinding wheels are specified by the type of abrasive, grit size, grade, structure and bond (*Figure 8.4*).

Aluminium oxide and silicon carbide wheels are the abrasives most widely used in general engineering. Silicon carbide is the harder material, but suffers chemical wear when cutting steels. It is therefore used for grinding cast iron and non-ferrous materials. The softer but tougher aluminium oxide abrasive is used for grinding steels. For more specialist applications, superabrasives have been developed. These are natural and synthetic diamond and cubic boron nitride (CBN). Due to their high cost, superabrasives are used mainly on materials which are difficult to grind; typically tungsten carbide, glass and ceramics.

8.3.1 Grit size

The abrasive grains making up the wheel are accurately graded and a wheel will usually contain only one size of grit. Wheels having a coarse grit size are used for roughing operations where rapid metal removal is required and surface finish is not critical. Finer grit sizes are used for finishing operations, giving superior surface finish but with a reduction in metal removal rate. The relationship betwen the type of work and the grit size is shown in *Table 8.2*.



*Optional symbols
The symbols 0 and 6 are the manufacturer's own

Figure 8.4 British Standard system for marking abrasive wheels (BS 4481: 1969: Part 1)

Table 8.2 Grit sizes used for different types of work

Type of work	*Grit size*
Roughing	8–24
General purpose	30–60
Fine finish	70–180
Extra-fine finish	220–600

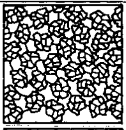
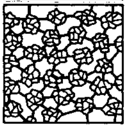
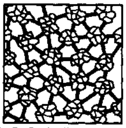
8.3.2 Wheel grade

The wheel grade indicates the strength of bond holding the wheel together. If the bond is weak the wheel is said to be 'soft'. When grits become blunt in a soft wheel they are readily torn out to expose new sharp grits; a mechanism known as 'self-sharpening'. Since soft wheels are less likely to dull, they tend to give a cooler cut. Hard wheels are strongly bonded together and are less likely to exhibit self-sharpening; their use is favoured in finishing operations. Wheel grade is specified by a code letter (A, soft; to Z, very hard); the complete range of letters is shown in *Table 8.3*.

8.3.3 Wheel structure

The wheel structure refers to the overall arrangement of the abrasive grit, bond and void within the wheel. An open structure with a large amount of void would be preferred for roughing operations involving high rates of metal removal, since an open structure is less likely to load and has good self-sharpening characteristics. Wheels having a dense structure are favoured for finishing operations where a good surface finish is required but, as dense wheels are more likely to load, only low rates of stock removal can be used. Dense wheels are less likely to self-sharpen. Structure is indicated by a numerical scale ranging from 0 (dense) to 12 (open) as indicated in *Table 8.4*.

Table 8.4 Numerical scale used to indicate wheel structure

Structure		*Symbol*	*Applications*
Dense		0 to 4	Finishing operations
Medium		5 to 7	General operations
Open		8 to 14	High rates of metal removal

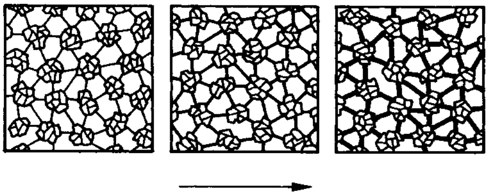


Figure 8.5 Effects of varying the amount of bond on the structure of a grinding wheel. (Courtesy of Carborundum Abrasives GB Ltd)

8.3.4 Bond

The type and quantity of bonding material used in the manufacture of a grinding wheel has a marked effect on its properties and performance. Strength, elastic modulus and flexibility are controlled by the bond, as is the ease with which grits can be dislodged by bond post-rupture (self-sharpening). *Figure 8.5* shows the effects of varying the proportion of bond on wheel structure.

For general applications, the most commonly used bonds are vitrified and resin. Less widely used bonds include shellac, rubber, oxychloride and metal.

Vitrified bonded wheels are used very extensively; this glass-type bond is strong, porous and unaffected by water, acids, oils or normal operating temperatures.

The resinoid bond is basically organic and wheels with this type of bond are more flexible than vitrified wheels. Resinoid wheels give high rates of stock removal and are widely used in foundries and fabrication shops for fettling and cutting-off operations. Shellac bonded wheels are capable of producing fine finishes on camshafts and mill rolls, but are unsuitable for heavy duty applications. Rubber bonded wheels are used for most feed wheels in centreless grinding and for specialised cut-off wheels which are particularly narrow ($\leqslant$1 mm).

Metal bonding is used mainly with superabrasives such as diamond and cubic boron nitride. The abrasive is bonded to a metal disc using powder metallurgy techniques or electroplating. Finally, the wheel-maker can include his own identifying mark. Superabrasives are specified using a similar system to that used for conventional materials (*Figure 8.6*).

Table 8.3 Code letters used to specify wheel grades

Wheel grade	*Symbol*
Soft	A to I
Medium	J to Q
Hard	R to Z

8.4 Mounting the grinding wheel

The Abrasive Wheel Regulations 1970 require that 'no person shall mount an abrasive wheel unless he has been trained to do so or is competent to do so'. Employers are also required to keep a register of persons appointed to mount grinding wheels and the class of wheels which they can mount.

Having selected the correct type and size of grinding wheel for the particular grinding machine or operation, the wheel should be examined carefully to ensure that it is free from defects. Visual inspection and a 'ring test' can find most defects. If the wheel is satisfactory it can be fitted to the

Superabrasive Marking System

Diamond Product Grading

MD	126	R	50	B	25	3 mm
ABRASIVE TYPE	*FEPA GRIT SIZE*	*GRADE*	*CONCENTRATION*	*BOND*	*SUFFIX*	*COAT DEPTH*
D	502 (Coarse)	I (Soft)	25 (Low)	B – Resinoid	Various	in millimetres
MD	to	to	to	M – Metal		
CMD	46 (Fine)	U (Hard)	100 (High)			

CBN Product Grading

CB	181	150	V	ST-H10
ABRASIVE TYPE	*FEPA GRIT SIZE*	*CONCENTRATION*	*BOND*	*SUFFIX*
XCB	252 (Coarse)	25 (Low)	B – Resinoid	Various
CB	to	to	M – Metal	
	46 (Fine)	200 (High)	V – Vitrified	

CBN – Cubic Boron Nitride

Figure 8.6 Superabrasive marking system. (Courtesy Carborundum Abrasives GB Ltd)

machine after a final check of wheel size, bore fitting, flanges and maximum operating speed. The wheel should fit easily onto the spindle and is clamped into position using the correct flanges and washers, as appropriate. Where a nut is used to clamp the wheel in position this should tighten up in the opposite direction of rotation of the wheel. Although the wheel should be held securely, over-tightening must be avoided. Where a torque figure is recommended, a torque spanner can be used.

8.5 Balancing and dressing

To ensure safe operation and to achieve the best quality of surface on the work piece, the wheel must be balanced and dressed. Several methods of balancing are available: static, dynamic and automatic balancing/self-balancing. In many applications, static balancing (which is simple and cheap) is quite adequate. With static balancing the balance weights are usually incorporated in the wheel flanges. Initially, the balance weights are set opposite each other, the wheel assembly is fitted to the grinding machine and then trued by passing a dressing diamond across its face. The wheel is now true but will be out of balance to some degree. To remove this 'out of balance', the wheel assembly is taken from the machine and mounted on a balancing arbor (*Figure 8.7*). The balance weights on the wheel flange are checked to ensure that they are positioned opposite each other (*Figure 8.8*(*a*)) and the wheel is placed on the balancing fixture. This fixture provides two parallel edges which can be levelled accurately. The wheel is allowed to rotate freely until it stops; the heaviest point is now at the bottom. Adjustments are now made to the balance weights which are moved until they are positioned as shown in *Figure 8.8*(*c*)). The balance weights are now moved a little at a time in the direction of the arrows (*Figure 8.8*(*d*)). After each small movement the wheel is allowed to rotate to see if it still stops with the marked position at the top. When the wheel is correctly balanced the mark will stop in a random position when the wheel is gently rotated. The wheel is now statically balanced and is replaced on the machine, dressed ready for grinding.

Dynamic balancing is performed with the wheel in place on the machine in a similar manner to wheel balancing on a car. Since this method is superior to static balancing it is used when machining critical components. Automatic balancing or self-balancing methods have been developed and have the advantage of correcting any imbalance which develops as the wheel wears.

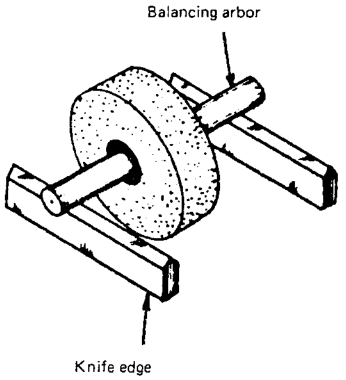


Figure 8.7 Grinding wheel mounted on an arbor ready for balancing

To prepare the wheel for use it must be dressed, an operation which ensures that the wheel is true and that the abrasive grits are in the correct condition for cutting. A profile can be dressed onto the wheel for manufacturing products such as threads and gears.

Dressing can be carried out in several ways, the most widely used methods being crush, diamond and roller dressing. Crush dressing uses a star wheel to ensure that the grinding wheel is clean, sharp and concentric on simple bench grinders and is unsuitable for precision grinding. For general dressing operations, diamond dressers are probably the most popular. Dressing is carried out by tracking the sharp dressing diamond across the face of the wheel which is rotated at normal speed. Shallow cuts are taken, removing a small layer from the surface of the wheel; loading and dulled grits are removed and concentricity maintained. Usually, several passes of the diamond are made progressively, reducing the depth of cut until the final pass may be at 'spark out'. Severity of dressing is one of the factors controlling wheel performance. Coarse dressing with a deep cut and high cross traverse speed gives a wheel with an open structure and well spaced sharp grits. A coarse dressed wheel is less likely to load, gives low cutting forces (the wheel cuts cool), high wheel wear and poor surface finish. Fine dressing where the diamond takes a small cut and is traversed slowly across the face of the wheel gives a wheel with more cutting edges, but these are blunt. Used in this condition the wheel will generate high cutting forces. Therefore, only

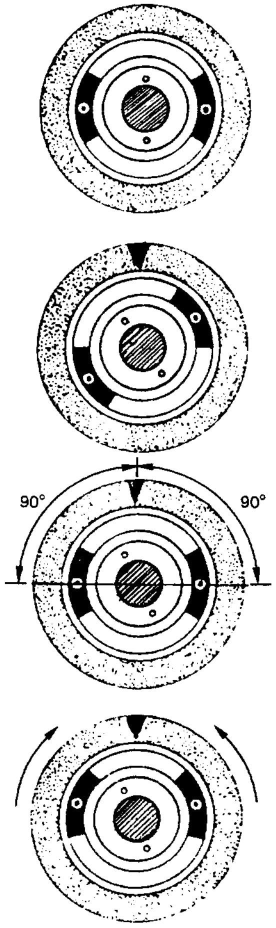


Figure 8.8 Procedure for balancing a grinding wheel using weights fitted to the wheel flange

shallow cuts can be made without the risk of thermal damage to the work piece. However, the larger number of grits and their relatively blunt shape produces a good surface finish. A summary of the effects of dressing on grinding performance is given in *Table 8.5*. If a profiled wheel is required, for example for the manufacture of gears or screw threads, this can also be produced by diamond dressing. The diamond is tracked across the face of the wheel using a proprietary system such as the diaform, special dressing attachment, or under computer numerical control (CNC). Crush dressing involves forcing the grinding wheel against a profiled dresser and crushing the grains, so that eventually the wheel takes up the shape of the dresser. The method is particularly suited to deep profiles and gives a sharp free cutting face to the wheel. The technique is often used for one-step manufacture of components for the automobile and aerospace industries by plunge or creep feed grinding.

The dressing/truing of superabrasives is minimised due to their hardness and cost. Since for most applications the abrasive is only a thin surface layer, it is best to take extra time in fitting the wheel to ensure the best concentricity. The wheel is then finally trued using an abrasive wheel dresser, removing the minimum of superabrasive material. Although truing and dressing are often seen to be overlapping in conventional grinding, this is not true with superabrasives. Dressing conditions the wheel by removing bond and exposing the grits at the surface; a dressing stick of aluminium oxide is used for this purpose, the wheel being rapidly rotated. When properly carried out, dressing should only expose the superabrasive grains and make no alteration in wheel geometry.

Table 8.5 Summary of the effects of dressing on grinding performance

Property	*Coarse dressing* (deep cut high cross-feed/rev.)	*Fine dressing* (shallow cut or spark-out small cross-feed/rev.)
No. of active grits	Small number	Larger number
Structure	Open	Close
Shape of grits	Sharp	Blunt
Cutting forces	Low	High
Risk of thermal damage	Low	High
Rate of stock removal	High	Low
Surface finish	Poor	Good

8.6 Grinding mechanics

The mechanism of metal removal in abrasive machining operations is highly complex. Several factors distinguish grinding from other metal cutting operations; the abrasive grits in the wheel are very small and randomly orientated; the ratio of tangential to normal force is 0.5 compared with a ratio of 2 in single-point cutting; and the specific energy u is high, typically 50 J mm^{-3} compared with 3 J mm^{-3} for single-point cutting. These observations suggest that grinding processes are significantly different from other metal cutting operations. Examination of the grinding debris (*Figure 8.9*) shows some chips of conventional appearance plus rather more randomly shaped chips, metal globules and abrasive particles. It is only the outermost grits which interact with the work piece by cutting, ploughing or rubbing.

Grinding theory is discussed here using the following symbols.

- F_n normal force on the work (N)
- F_t tangential force on the work (N)
- F_g average force per grit (N)
- D wheel diameter (mm)
- D_w work piece diameter (mm)
- d wheel depth of cut (mm)
- b width of cut or cross-feed (mm)
- V wheel speed (m min^{-1})
- v work speed (m min^{-1})
- μ grinding coefficient F_t/F_n
- u specific energy of grinding (J mm^{-3})
- t grit depth of cut (mm)
- G grinding ratio (volume of metal removed/volume of wheel removed)
- C number of active grits per unit area (grits/mm^2)
- r width/depth ratio of an average groove (equal to half the angle of an average groove)
- Δ_a average slope of the ground surface

Any attempt to model the grinding process must of necessity involve a number of assumptions. From the earliest studies of Alden[2] and Guest,[3] grinding has been regarded as a form of micromilling, the abrasive grits acting as randomly orientated cutters. One of the most widely used parameters derived from the modelling procedures is the concept of grit depth of cut t. This is the depth of cut taken by an average grit (*Figure 8.10*). Backer *et al.*[4] have derived the following widely used expressions for grit depth of cut in surface and cylindrical grinding:

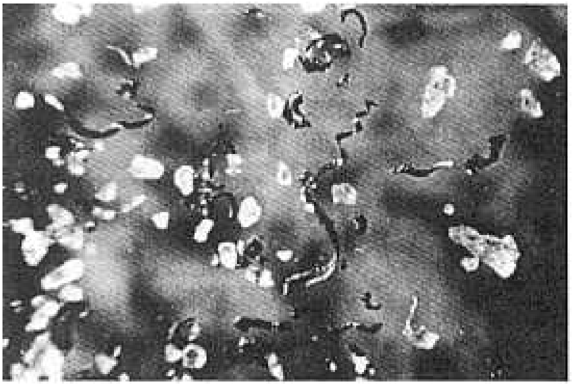


Figure 8.9 Grinding debris showing chips, metal globules and abrasive particles

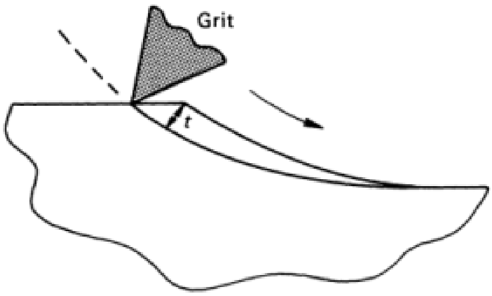


Figure 8.10 Diagram showing the grit depth of cut concept

$$\text{Surface grinding } t = \left[\frac{4v}{VCr}\sqrt{\frac{d}{D}}\right]^{1/2} \quad (8.3)$$

$$\text{Cylindrical grinding } t = \left[\frac{4v}{VCr}\sqrt{\left(\frac{d(1 + D/D_w)}{D}\right)}\right]^{1/2} \quad (8.4)$$

Equation (8.3) for surface grinding is a unique form of equation (8.4) with $D_w = \infty$.

All the functions in the above equations are readily available with the exception of the number of active grits per unit area C and the width/depth ratio r. One method of determining the value of C is to roll the wheel across a piece of smoked glass and then count the number of points where the glass is clear. This method has several disadvantages: the wheel must be removed from the machine to take a reading and the method only counts the outermost grits, giving no indication of any changes which may occur as the wheel penetrates the work piece. An alternative method uses a surface measuring instrument fitted with a chisel-edged stylus capable of giving a high-spot count p_c. The number of active grits per unit area is given by the high-spot count divided by the area traversed (width of stylus × traverse length). Sophisticated surface measuring equipment can indicate the high-spot count at varying levels, thus allowing the value of C to be determined at varying levels below the outer surface of the wheel. The value of C depends on the properties of the wheel, in particular the grit number, the dressing procedure and the grit depth of cut. Using the stylus method, the value of C can be monitored whilst the grinding wheel is mounted on the machine. In their original work, Backer *et al.*[4] determined the width/depth ratio r of an average scratch by taking a taper section of the work piece surface and measuring a number of typical scratches. More recently,[5] it has been suggested that the average slope Δ_a of the work piece surface could be used to derive the value of the width/depth ratio r. The value of r can be obtained using the expression:

$$r = 2 \tan (90 - \Delta_a) \quad (8.5)$$

The average slope of the work piece is measured using a stylus surface-measuring instrument of the appropriate type.

Taper sectioning is a destructive method, whereas stylus techniques are non-destructive and allow the process to be monitored continuously.

Although stylus methods can be carried out without removing either the wheel or the work piece from the machine and the values for C and r are more reproducible, the values obtained are only averages. Consequently, the value calculated for the grit depth of cut t will also be an average.

Once values have been determined for C and r, the grit depth of cut t can be calculated by substituting the grinding conditions in equation (8.3) for surface grinding or in equation (8.4) for cylindrical grinding.

For a surface grinding operation carried out under the following conditions

work speed v = 0.5 m s^{-1}
wheel speed V = 30 m s^{-1}
depth of cut d = 0.04 mm
wheel diameter D = 200 mm
No. of grits C = 2 grits/mm^2
width/depth ratio r = 18

the value of the grit depth of cut is calculated by substituting in equation (8.3):

$$t = \left[\frac{4 \times 0.5}{30 \times 2 \times 18}\sqrt{\frac{0.04}{200}}\right]^{1/2}$$

$$= 0.005 \text{ mm}$$

This calculation shows that the grit depth of cut is very much less than the wheel depth of cut, typically only about one-tenth. Despite the fact that it is difficult to determine a reliable value for the grit depth of cut t, the concept is very useful for predicting the effects of changes in grinding conditions on grinding performance.

The ability of a wheel to self-sharpen will depend to some extent on the average load per grit F_g. Assuming the average force per grit is proportional to the cross-sectional area of the undeformed chip, equation (8.3) can be rearranged to show that:

$$F_g \propto \frac{v}{VC}\sqrt{\frac{d}{D}} \qquad (8.6)$$

As the value of F_g increases, the wheel will be more likely to self-sharpen and the wheel will thus act softer. Conversely, reducing the average load per grit will make the wheel act harder as it will be less likely to self-sharpen. The average load per grit F_g will increase as the work speed v and wheel depth of cut d are increased, and F_g is reduced as the wheel speed V, the number of active grits C and the wheel diameter D are increased. Dressing can affect the average load per grit in two ways: an open wheel will only have a few active grits and, therefore, a high average load per grit; and sharp grits will carry less load than blunt grits. The surface finish R_a produced in a grinding operation can be estimated from the grit depth of cut t using the equation

$$R_a \approx 0.25t \qquad (8.7)$$

This equation assumes that the depth of an average groove on the work piece is equal to the grit depth of cut. By reference to equation (8.3) the factors favouring a good surface finish can be identified as those which give a small value for t. Thus in order to produce a good surface finish a low rate of metal removal (small values for v and d), a finely dressed wheel with a small grit size (large values of C and r) and a large diameter wheel (D) are required. In practice, grit depth of cut is not the only factor influencing surface finish; machine condition, work piece material, system stiffness and coolant can also have a significant effect on surface finish.

The mechanism of metal removal in the grinding process is complex and the grinding forces arise as a result of cutting, ploughing and rubbing. Indeed, as a freshly dressed wheel is used, wear flats form rapidly on the grits, increasing forces without contributing to the cutting processes. If, as in precision grinding, there is little self-sharpening, the blunting of the grits due to wear flats will control the frequency of dressing.

Grisbrook's[6] classic work on surface grinding shows the progressive increase in grinding forces with the number of passes of the work piece under the wheel (*Figure 8.11*). Hahn[7] has suggested that in controlled force grinding processes, the

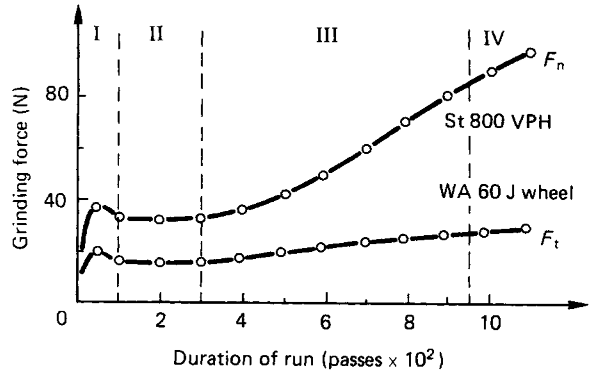


Figure 8.11 Relationship between grinding forces (normal F_n and tangential F_t) and duration of run

wheel develops optimum sharpness; when too sharp the wheel will blunt and when the wheel is dull it will self-sharpen (*Figure 8.12*).

The magnitude of the forces developed in grinding will depend on the machine settings, the work piece material and the condition of the wheel (dull or sharp). Grinding forces affect many aspects of the processes, including power consumption, flash temperatures, grinding ratio, specific energy and the risk of work piece damage.

The energy input to the processes is used in three ways: cutting, ploughing and rubbing. The specific energy u is, therefore, made up of three terms:

$$u = u_{\text{cutting}} + u_{\text{ploughing}} + u_{\text{rubbing}} \qquad (8.8)$$

The overall value for the specific energy is given by

$$u = \frac{VF_t}{vbd} \qquad (8.9)$$

Much of the energy expended in the grinding process is dissipated as heat, which is divided between the chip, abrasive grit and the work piece. Thus very high flash temperatures are

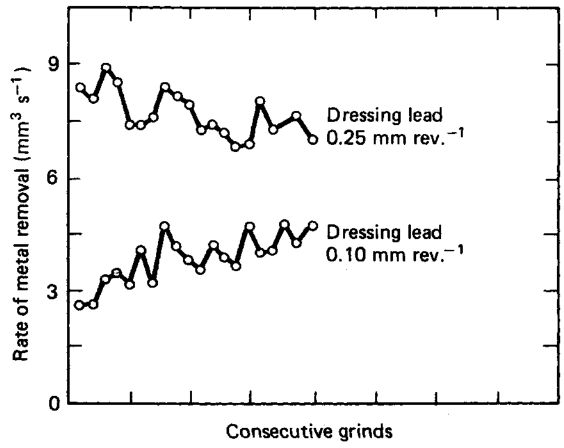


Figure 8.12 Effect of dressing lead on metal removal rate as grinding progresses

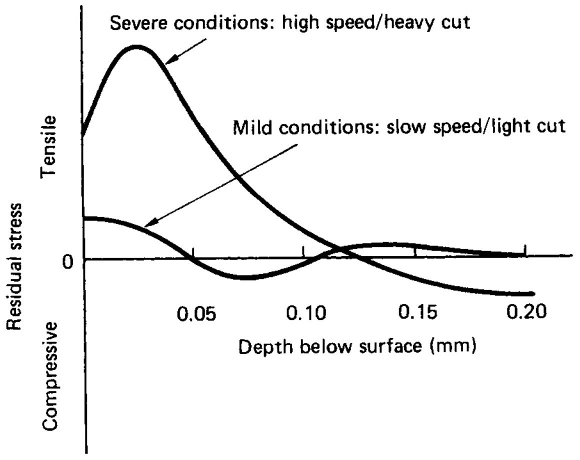


Figure 8.13 Effect of grinding conditions on residual stresses

generated at the chip–grit interface which will influence the wear of the grit. A proportion of the heat is carried away by the chip and the remainder goes into the work piece. Even when coolant is present the immediate surface of the work piece may be heated to temperatures which will produce changes in both structure and properties. Hardened steels are particularly prone to grinding damage. This can take the form of tempering where soft spots are produced on the surface of the component or, if conditions are very severe, the surface may be raised to the hardening temperature of the steel which on rapid cooling will produce martensite. Residual stresses are also produced as a result mainly of the temperature gradients associated with grinding. Typical stress distributions in the surface of ground components are shown in *Figure 8.13*. If conditions are exceptionally severe, the stresses may produce grinding cracks; these are generally at right angles to the direction of grinding. To avoid grinding damage a coolant should be used and the grinding conditions should not be too aggressive. Since grinding damage is often difficult to detect, critical components should be inspected.

Grinding cracks in steel components can be readily found using magnetic crack detection. Tempering and martensite formation are best detected by lightly etching the steel in the appropriate etching reagent. Tempered areas show up darker than the surroundings and martensite, burnt or glazed areas are lighter than the surroundings.

8.7 Wheel wear

Wheel wear is both an important and a complex parameter. Both the grinding ratio and the frequency of dressing are related to the rate at which the abrasive wears. Three major mechanisms of abrasive wear are: attritious wear, fracture and bond post-rupture.

8.7.1 Attritious wear

With this type of wear the abrasive grain is dulled due to a wear flat forming by a process called 'attrition'. The grit and the work piece react at the high flash temperatures associated with grinding, with wear taking place by a combination of physical and chemical processes. Attritious wear is lower between materials which are chemically inert; aluminium does not react strongly with steel and so gives a lower rate of attritious wear than silicon carbide which will dissolve in molten iron.

8.7.2 Fracture

The grains of abrasive are hard and brittle so that if the forces on the grit build up as a result of wear-flat formation the grit may fracture and expose fresh cutting facets. The ease with which a grit fractures is called its 'friability'; good friability in an abrasive means that it is easily fractured. Aluminium oxide is less friable than silicon carbide and the grits in an aluminium oxide wheel tend to self-sharpen at a slower rate than those of silicon carbide.

8.7.3 Bond post-rupture

When a grit has become very dull and worn, the cutting forces may be sufficient to fracture the bond removing the grit and thus expose fresh abrasive. The ease with which a grit can be removed depends on the grade of wheel; soft wheels will self-sharpen easily maintaining a continuous supply of sharp grits and hence give low cutting forces, reducing the risk of thermal damage and residual stresses.

It must also be remembered that in many grinding operations it is the dressing operation which accounts for the bulk of wheel wear.

8.8 Grinding ratio

Grinding is no different to other metal cutting operations in that, as material is removed from the work piece, there is corresponding tool wear; in this case wheel wear. The grinding ratio G is a method of expressing this wear and is calculated using:

$$G = \frac{\text{Volume of metal removed}}{\text{Volume of wheel removed}} \tag{8.10}$$

Many factors affect the grinding ratio, including work piece material, type of wheel, and machine conditions. Although there are some advantages in achieving a large value for G (typically 100–500 in fine grinding), such as a longer wheel life, it may lead to duller grits, higher forces and an increased possibility of damage to the work piece. The rate of wheel wear G is not constant throughout the grinding cycle. *Figure 8.14* shows volume of metal removed plotted against the volume of wheel removed. It shows that, initially, the freshly dressed wheel wears rapidly, but the rate of wear declines until a steady-state region is reached. It is in this region that the bulk of grinding takes place. Normally the wheel is redressed before the third stage where the wear rate begins to accelerate.

8.9 Grinding forces

Some estimate of grinding forces can be made by relating them to the specific energy. Equation (8.11) relates the power used in surface grinding to the tangential grinding force F_t, machine settings and specific energy u:

$$\text{Power} = F_t V = bdvu \tag{8.11}$$

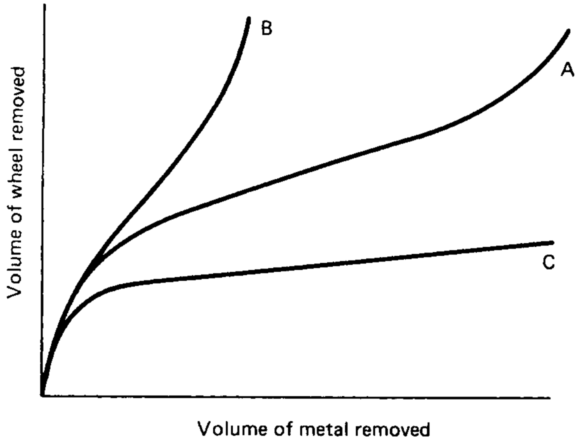


Figure 8.14 Volume of metal removed vs. volume of wheel removed; A, wheel satisfactory; B, wheel too soft; C, wheel too hard

hence

$$F_t = \frac{bdvu}{V} \tag{8.12}$$

Unfortunately, the specific energy is not a constant even for one material as it depends on the condition of the wheel, the grinding conditions and the grit depth of cut. However, if an appropriate value is assumed for u, both F_t and the power can be estimated.

Similarly, the normal force F_n can be estimated using the expression for the grinding coefficient and rearranging:

$$\mu = \frac{F_t}{F_n} \tag{8.13}$$

hence

$$F_n = \frac{F_t}{\mu} \tag{8.14}$$

Estimated values of the forces in grinding are of considerable value since their practical determination requires very specialist equipment.

8.10 Coolant

Although grinding can be carried out dry in most applications, there are considerable advantages in using a coolant or lubricant. The most widely used coolants are water based soluble oils. Since the coolant is continuously recycled it must undergo a reconditioning treatment before re-use. The contaminants in the coolant (worn abrasive, and metallic and oxide particles) can be removed in several ways. The simplest and most economical method consists of a series of holding tanks and weirs; the grinding debris settles out by gravity. An alternative system is to filter the coolant through a continuously fed filter paper, and for very critical work this method can be enhanced by means of a centrifuge.

Having ensured an adequate supply of clean coolant, it is essential that it is applied correctly. Grinding damage can occur if coolant supply is insufficient and this damage is not always obvious, hence the need to inspect all critical parts after grinding.

8.11 Grinding processes

Grinding can be carried out on a wide variety of materials using a range of configurations. Conventional processes consist of surface grinding, cylindrical grinding (internal and external) and centreless grinding. Machines are designed to traverse the work past the wheel using cross-feed; alternatively, the work is fed directly into the wheel—a technique known as 'plunge grinding'. By dressing a profile onto the wheel, complex shapes such as screw threads and gear teeth can be produced.

Grinding is a relatively expensive method of machining and is generally used as a finishing process where good surface finish and accurate dimensions are required. Grinding is also applied to materials which are difficult to machine by conventional methods, e.g. hardened steels, aerospace materials, glass and ceramics.

8.11.1 Surface grinding

Surface grinding machines use a range of configurations, the most common being a reciprocating table and a horizontal spindle. The work reciprocates under the wheel, and most machines provide automatic cross-feed and automatic downfeed. There is also increasing demand for variable wheel speed. Components can also be plunge ground. Complex shapes can be manufactured by dressing a profile on the wheel using the appropriate attachment. Creep feed grinding is usually carried out in the surface-grinding mode. This process differs from conventional grinding in that the wheel depth of cut d is increased considerably (approximately 1000) and the work speed v reduced proportionally; hence the term 'creep-feed grinding'. Machines used for this type of grinding operation are normally special purpose if the process parameters are to be adequately controlled. Control of coolant application and wheel wear are particularly critical; often the wheel is continuously dressed throughout the process, appropriate adjustments being made for changes in wheel diameter. Creep-feed grinding is used to machine deep complex profiles into the work piece; typical products being the fir-tree serrations of gas turbine blades and connecting rod assemblies (*Figure 8.15*).

Surface grinding machines are available with computer numerical control (CNC), allowing the complete process to be carried out under computer control, including the dressing cycle, machining and gauging. In surface grinding CNC gives many benefits, including reduced cycle times, complete automation of the grinding cycle, more consistant quality and improved productivity. Further improvements in machine utilisation can be achieved by preparing programs off-line.

The complexity of product which can be produced by CNC grinding is shown in *Figure 8.16*(*a*). In this example, the work piece is a semiconductor die used for the manufacture of integrated circuits; eight components, ganged in line, were machined at the same time. In each component there are 20 slots, each 1.5 mm wide by 2.5 mm deep. The slots are at 2.54 mm pitch, and the total cumulative pitch error between the first and last slot is limited to ± 30 μm.

The total process is fully automated, and starts with CNC wheel dressing to produce the profile shown in *Figure 8.16*(*b*). Once the wheel has been formed the grinding cycle then deals with each slot in turn, progressively taking each slot to full

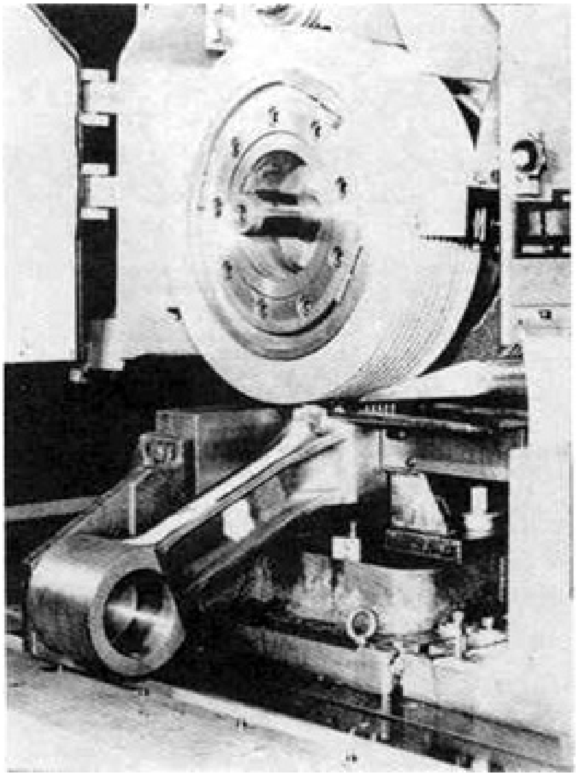


Figure 8.15 A connecting rod set up for creep-feed grinding using a formed wheel on an Elb grinding machine. (From Marsh[8])

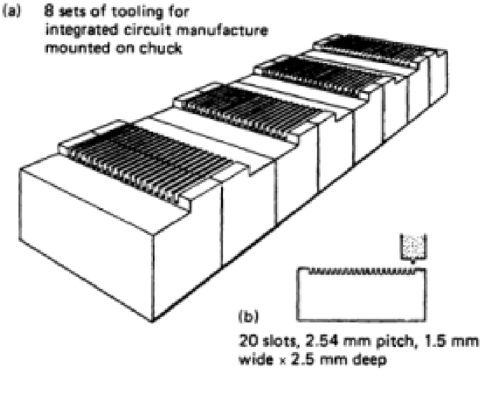


Figure 8.16 (a) Eight sets of tooling for integrated circuit manufacture, ganged in-line. (b) Wheel profiled ready for grinding. (Courtesy of Jones and Shipman plc)

depth from the solid, and with 10 in-cycle dressings applied as an integral part of the programme.

There is automatic size compensation for reduction in wheel diameter, and automatic increase in wheel speed to maintain a constant peripheral speed of 30 m s^{-1}.

An increment in down-feed is automatically applied at each reversal of the table, and the total down-feed required to bring each slot to full depth is divided (in the programme) into coarse, medium and fine sections. Finally, there is a 15-stroke spark-out before the wheel head is automatically raised and the table indexed transversely to position it in readiness for grinding the next slot.

The total time to CNC grind the eight components is substantially less than that needed by skilled workers using manually operated surface grinders.

Gains resulting from CNC grinding are improved productivity, total automation of the process, improved quality and a reduction in batch-to-batch variation.

Vertical spindle reciprocating table (*Figure 8.17(a)*) and vertical spindle rotary table (*Figure 8.17(b)*) grinding machines are used to produce flat surfaces in quantity production applications. Rotary table horizontal spindle machines (*Figure 8.17(c)*) are capable of producing very flat surfaces, but are generally of limited capacity.

8.11.2 Work clamping

For surface grinding the work is usually held in position using magnetic, electrostatic or vacuum chucks. Obviously, magnetic chucks can only be used with ferromagnetic materials; electrostatic chucks will work on any electrically conducting material; and vacuum chucks will work on metallic and non-metallic materials. It is normally advisable to demagnetise components which have been held in a magnetic chuck.

Conventional jigs/fixtures are used for holding complex shapes requiring grinding.

8.11.3 Cylindrical grinding

Cylindrical grinding can be carried out on both external and internal cylindrical surfaces. External grinding provides the best grinding conditions with a small arc of contact (*Figure 8.18*), whereas internal grinding gives a larger arc of contact and, as a consequence, more severe conditions. External grinding is carried out between centres or in a chuck. The grinding wheel runs at normal speed (*c.* 30 m s^{-1}) and the work piece at 20–40 m min^{-1}. Wheel and work run in opposite directions at their point of contact. With cylindrical grinding there is often the facility to realign both wheel and table (universal grinding machine). Cylindrical grinding can be used to produce a wide range of products by traverse and plunge grinding with flat or profiled wheels. These products include straight cylinders, screw threads, tapers and complex shafts.

CNC has been applied to cylindrical grinding, giving similar benefits to those obtained when it is used with surface grinding. However, when grinding multidiameter shafts, additional advantages are obtained; in particular, all the operations are carried out without removing the component from the machine, thus ensuring concentricity of all diameters and giving very large reductions in floor-to-floor times.

A Jones and Shipman Ten[9] CNC cylindrical grinding machine is shown in *Figure 8.19* together with its control system. As an aid to programming the Series Ten[9] machine provides the operator with a series of coded ‘canned cycles’, examples of which are shown in *Figure 8.20*. Circular and linear interpolation are available for contour generation, both when dressing and when grinding (*Figure 8.21*). The grinding sequence for manufacturing a typical shaft is shown in *Figure 8.22*. Once the component has been placed in the machine, the grinding operations follow the sequence listed. In general, the sharpest curves/tapers are dressed onto the wheel, whilst other less sharp features are cut under CNC control by interpolation. In-process gauging allows adjustment of end stops and wheel speed to compensate for changes in wheel diameter. CNC grinding ensures that the machine spends a greater

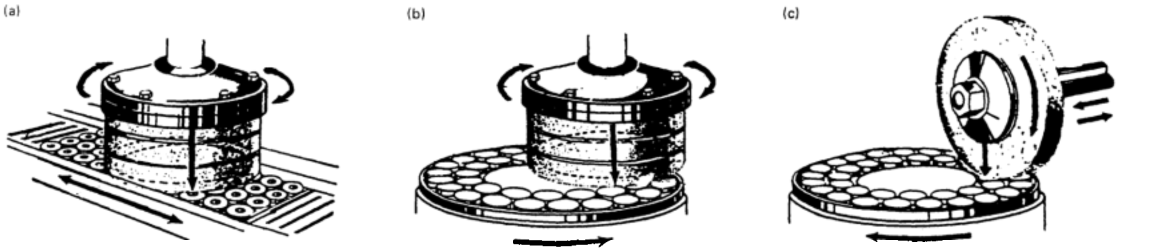


Figure 8.17 (a) Vertical spindle reciprocatingtable grinding machine. (b) Vertical spindle rotary table grinding machine. (c) Horizontal spindle rotary table grinding machine. (Courtesy of Carborundum Abrasives GB Ltd)

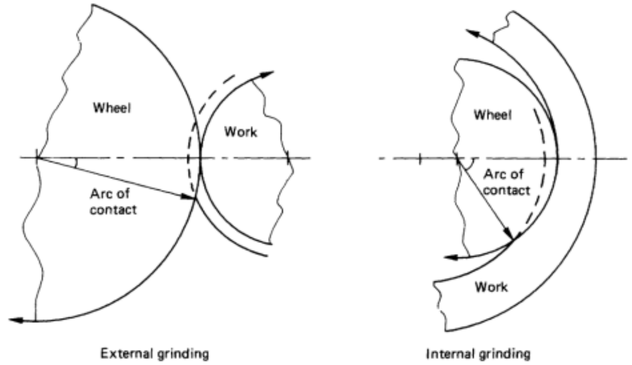


Figure 8.18 Effect of wheel–work configuration on arc of contact

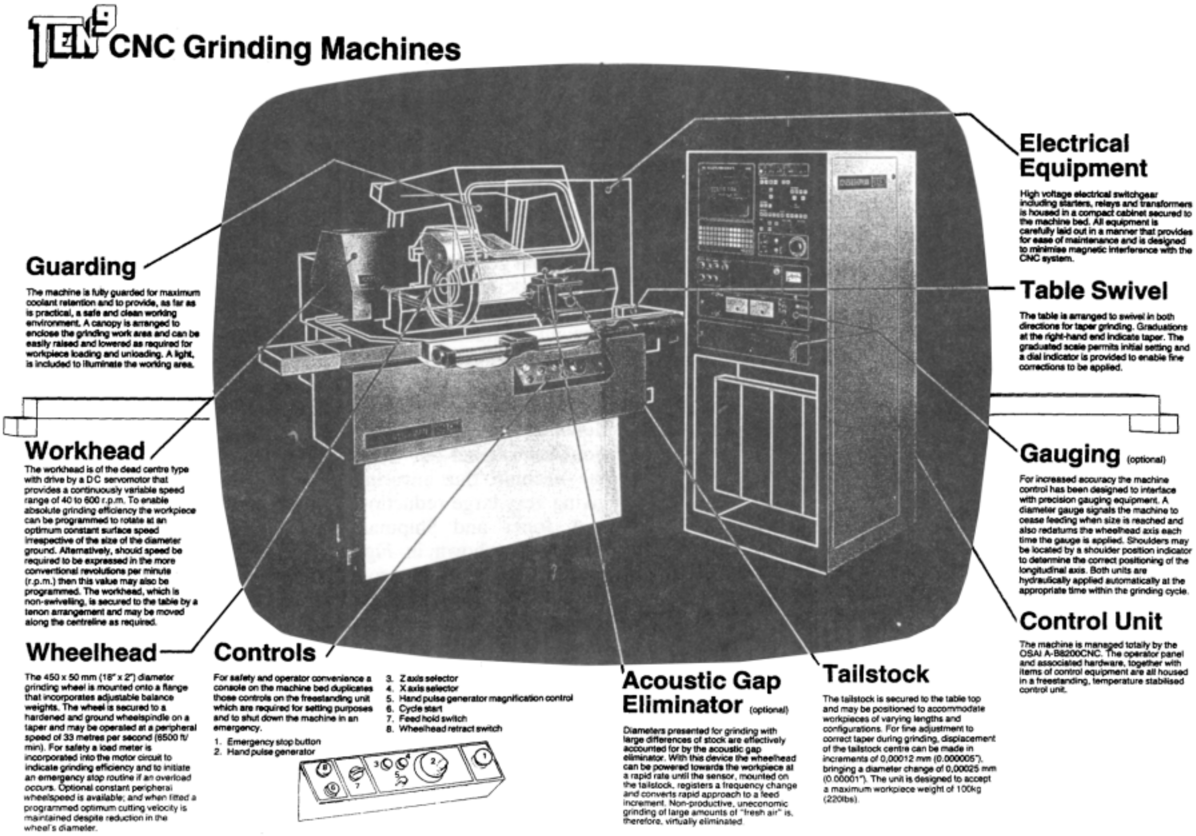


Figure 8.19 Jones and Shipman Ten^9 CNC grinding machine. (Courtesy of Jones and Shipman plc)

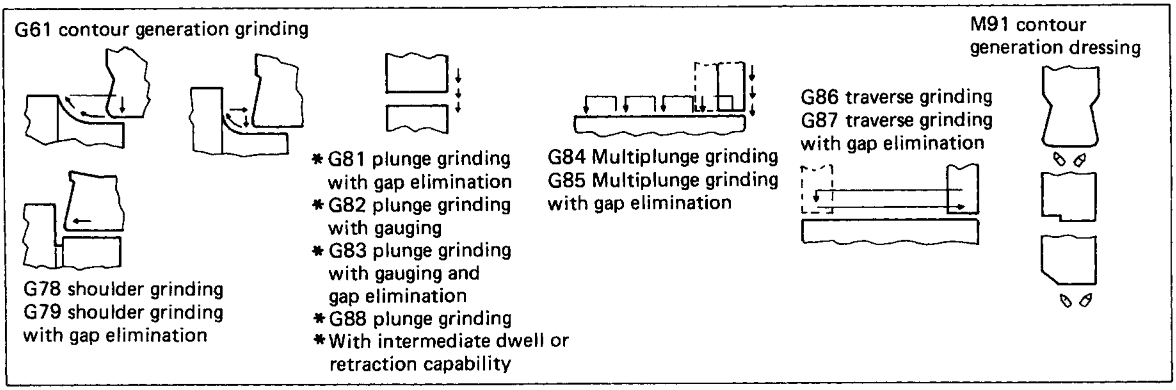


Figure 8.20 Examples of 'canned cycles' which enable more efficient programming. (Courtesy of Jones and Shipman plc)

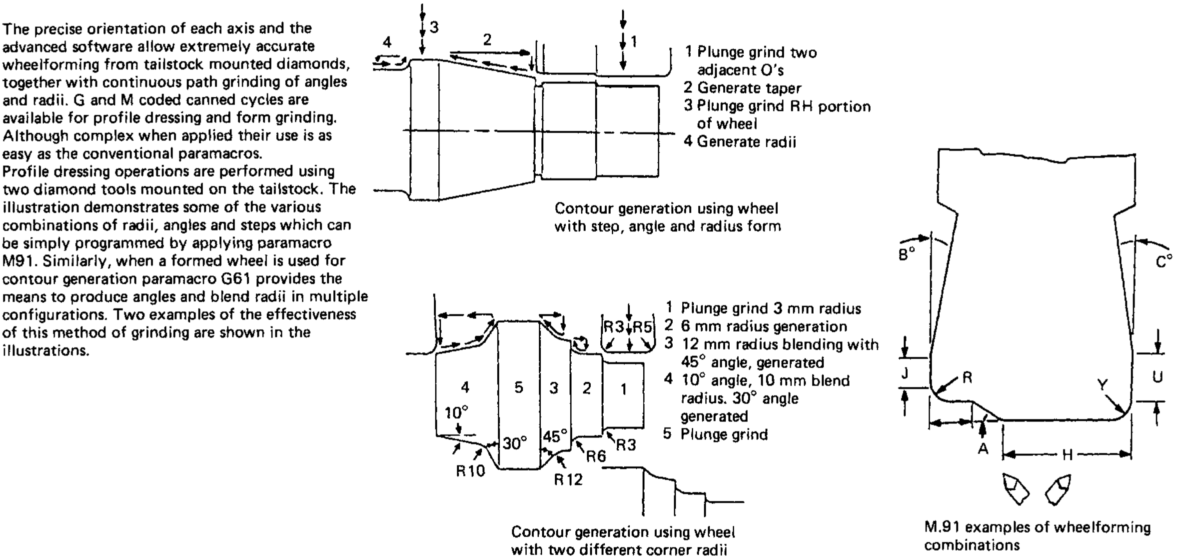


Figure 8.21 Circular and linear interpolation for contour generation. (Courtesy of Jones and Shipman plc)

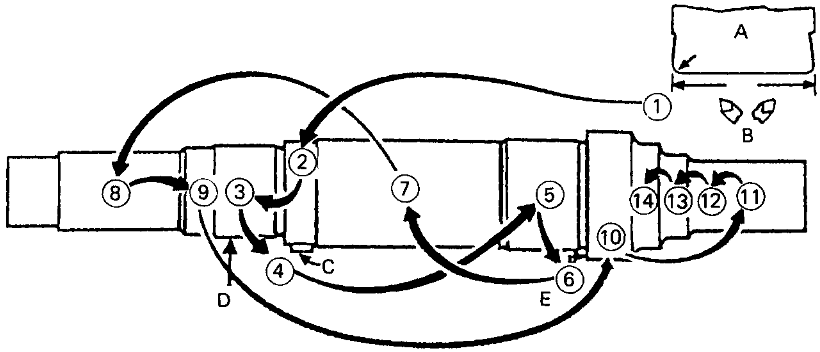
proportion of its time cutting metal and gives very substantial reductions in cycle times. Components need no longer be moved in batches from machine to machine and a 'just-in-time' (JIT) processing approach becomes possible.

Internal cylindrical grinding is used in the manufacture of precision products such as the races of ball and roller bearings. The wheel work configuration gives a long arc of contact and, as the wheel is often a close fit in the bore, the application of coolant and the removal of grinding debris can be difficult. An added problem is the need to run the wheel at very high speed (*c*. 30 000 rev. min^{-1}), due to its small diameter, in order to maintain a normal cutting speed of 30 m s^{-1}. Frequently, the grinding spindle (quill) is also of small diameter and thus the grinding forces can readily deflect the wheel, giving problems with system stiffness and parallelism when grinding. The problem of parallelism has been overcome by an adaptation of the process called 'controlled force grinding' or 'constant force grinding'. By maintaining a constant force the deflection of the quill will be constant and this deflection can be allowed for in the machine settings. Machines using this technique are widely used in the manufacture of ball and roller bearing tracks.

8.11.4 Centreless grinding

Centreless grinding can be readily automated and is, therefore, a widely used production process; both external and internal surfaces can be ground by this process. Since the work piece does not require mounting between centres or holding in a chuck, fast cycle times can be achieved.

The work–wheel arrangement in centreless grinding is shown in *Figure 8.23*. A work rest blade carries the work, which is ground betwen two wheels; the larger wheel being the grinding wheel and the smaller one being a regulating wheel which is rubber bonded. Both wheels are wider than those used in conventional grinding in order to reduce the effects of wheel wear. The regulating wheel rotates at a much slower speed (20–60 m min^{-1}) than the grinding wheel, and controls both the rotational and longitudinal motion of the work piece.



A Wheel B Diamonds C Steady D Diameter gauge
E Shoulder location gauge

Sequence to achieve EP7 class bearing shaft tolerances on diameters 3 and 5

1 Profile dressing of wheel
2 Plunge grind steady ϕ with gap elimination. (Program stop, apply steady)
3 Plunge grind with ϕ gauge
4 Shoulder grind with gap elimination
5 Plunge grind
6 Shoulder grind with gap elimination
7 Multiplunge grind
7 Traverse grind
8 Plunge grind
9 Plunge grind
10 Plunge grind
11 Plunge grind
12 Shoulder grind
13 Generate radius
14 Generate taper and blend radius

N	G	M	S/T	X	Z	Assigned/paramacro parameters
N10						(Spindle shaft)
N20						G71 G94 G97 G80 (CM.CAN)
N30		M36				
N40		M04	S100			(AP, P100 = 75)
N50		M60				
N60		M91		X20	Z145	B2 C2 D2 E.151.015 J5 P2 Q0
N70	G96		S15			
N80	G81			X56.900	Z145	A.015 P57.3 S10 Q0
N90		M00	S25			(Apply steady)
N100	G82			X50.000	Z166	A.015 P50.3 150.1 Q0
N110	G79				Z166	E.005 P50.02 S5 K-164.9 W-165.1
N120	G88			X60.000	Z1	A.015 P60.3 S5 Q0
N130	G79				Z1	E.005 P60.02 S5 K.1 W-.1
N140	G84			X57.000	Z45	A2 P57.3 S3 F1 K-150 Q1
N150	G86			X57.000	Z150	E.10 P57.004 S2 K45 Q100 R2 L2
N160	G88			X42.000	Z220	A.015 P42.3 S5 Q0
N170	G88			X49.800	Z197	A5 P50 S5 F1 Q0
N180	G88			X70.000	Z5	A.020 P70.3 S5 Q0
N190	G88			X38.000	Z55.3	A.015 P38.3 S5 Q1
N200	G78				Z56	E.003 P38.005 S5 K55.2 W55.32
N210	G61			X44.000	Z50	C.015 D.01 E1 P44.3 K40 R6 U56
N220	G61			X56.000	Z35	B60 C.015 D.01 E1 P56.3 K25 R4 U70
N230		M60				
N240		M02				

Figure 8.22 Sequence to machine a typical shaft using a CNC cylindrical grinding machine. (Courtesy of Jones and Shipman plc)

The work is forced against the work rest, the forward feed rate f being controlled by slightly inclining the regulating wheel. An approximate value of the feed rate is given by:

$$f = D_r N \sin \gamma \quad (8.15)$$

where f is the feed rate (mm), D_r is the diameter of the regulating wheel (mm), N is the number of revolutions per minute of the regulating wheel, and γ is the angle of inclination of the regulating wheel.

External centreless grinding is particularly suited to the manufacture of cylindrical products of constant diameter by through feeding; almost any length can be produced and the

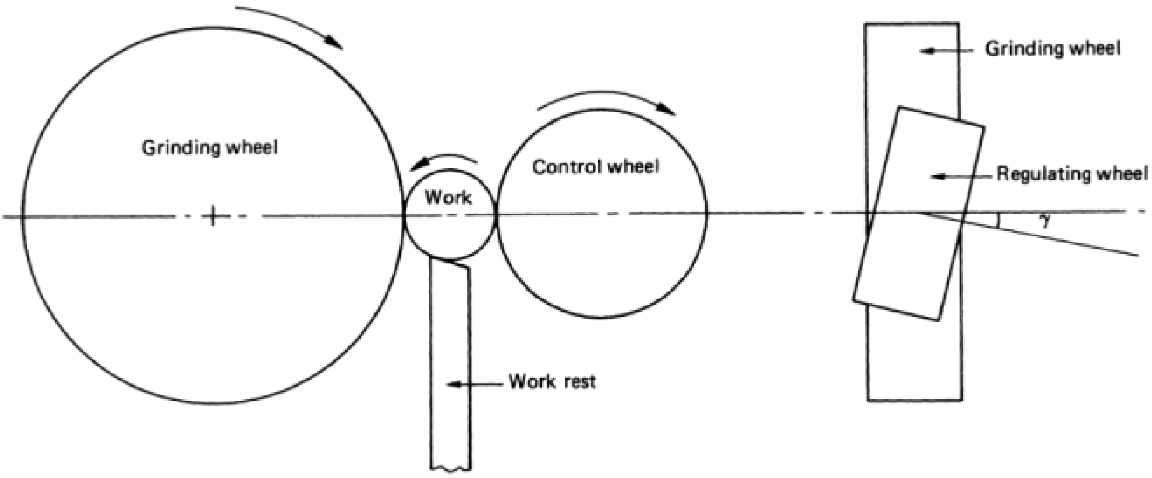


Figure 8.23 Wheel–work arrangement for centreless grinding

process is readily automated. However, the process can be configured to produce tapered or multidiameter cylindrical components. By profiling the wheel, spherical products can be made.

Internal centreless grinding is carried out by supporting the work piece between rollers and the grinding wheel moves into the work piece. To ensure that the internal and external diameters are concentric, the external aspect must be accurately machined before internal centreless grinding.

Centreless grinding processes are particularly suited to quantity production and are readily automated. However, there are restrictions on the shape of product which can be manufactured.

8.11.5 Tool and cutter grinding

Grinding is widely used in the manufacture and reconditioning of cutting tools. This work is usually carried out on universal tool and cutter grinders which are configured to grind complex tooling such as milling-cutters, single-point cutting tools, hobs and reamers. Specially shaped grinding wheels are normally used for this type of operation, and for the more exotic tool materials the use of superabrasives may be necessary.

8.11.6 Cutting off

Grinding is frequently used in fettling/snagging operations in the foundry to remove down runners and feeders from castings. Machines designed for this purpose achieve high metal removal rates using resinoid/rubber bonded wheels.

8.12 Newer abrasives and grinding techniques

Considerable research is being carried out on both abrasives and machining methods with the object of further improving grinding performance.

A new abrasive, known as SG, has been developed by Norton Abrasives. Chemically, the new material is aluminium oxide, but its unique structure gives it levels of performance similar to superabrasives at a price similar to conventional abrasives.

Slightly purer and about 10% harder than aluminium oxide used in the manufacture of ordinary grinding wheels, the new material is a ceramic produced by sintering the basic material produced by a seeded gel technique; hence the name SG. It is its structure which gives SG its enhanced properties. *Figure 8.24* shows an aluminium oxide wheel alongside a SG wheel of the same grain size. A traditional fused oxide grit is made up of a few, or possibly a single, grit. By contrast, the same single SG grain will contain millions of submicrometre particles. Each time a traditional aluminium oxide grit fractures in use, it can lose as much as one-fifth of its grinding surface, periodic dressing being needed to restore grinding performance. SG abrasive fractures on a much finer scale due to the submicrometre particle size of its constituents; cutting efficiency is maintained at significantly lower wear rates with SG wheels.

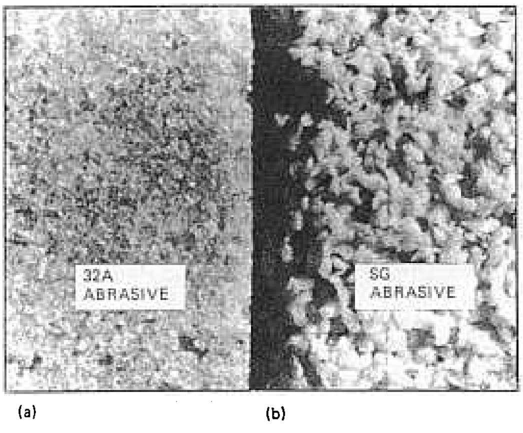


Figure 8.24 (a) Aluminium oxide wheel and (b) SG abrasive of the same grain size[9]

Table 8.6 Improvement in performance obtained by using SG grinding wheels[9]

Wheel diameter (mm)	*Wheel life increase* (%)			*Increase in grinding rate* (%)		
	Oil	*Water*	*Dry*	*Oil*	*Water*	*Dry*
< 350	310	363	167	220	97	650
> 350	440	340	nil	140	100	nil

Since the grits remain sharp, even with minimal dressing, the risk of metallurgical damage to the work piece is much reduced. It is claimed that vitrified SG wheels give longer wheel life, increased productivity, require less frequent dressing and reduce the risk of metallurgical damage.

A summary of grinding test results comparing the performance of SG grinding wheels with conventional aluminium oxide wheels is given in *Table 8.6*. SG abrasive wheels are suitable for both ferrous and non-ferrous materials up to a hardness of $67R_c$. Cost savings due to reduced cycle times are proportionally greater when using semiautomatic and CNC grinding machines.

It is claimed[9] that prime sectors for the application of SG abrasives are the aerospace industry and for critical components particularly susceptible to metallurgical damage.

One obvious method of making the grinding process more competitive is to increase metal removal rates, and this is the main objective of high-speed grinding. Such high-speed-grinding machines require very careful design, paying particular attention to guarding the grinding wheel, frequently by the use of armour plate materials. Special grinding wheels have been developed for use at high speeds (*c.* 60 m s^{-1}), and this has involved the building of special speed testing facilities.

Even with grinding there are some materials which can present machining difficulties. Electrochemical grinding was developed as a method of eliminating these problems. The process is carried out on a special-purpose machine, the bulk of the material being removed from the work piece by electrochemical action, the grinding wheel imparting surface finish and dimensional accuracy.

8.13 Honing

Honing should no longer be regarded as just a finishing process for cylindrical surfaces (internal and external). The process has been developed to the stage where it can frequently provide a cost-effective alternative to grinding and fine boring. Modern honing machines are capable of removing hard or soft materials at economical machining rates, whilst giving a product accurately to size, of good geometric shape (roundness and taper) and a controlled finish. Machines currently available range from manually operated machines to fully automated multispindle machines with microprocessor control.

8.13.1 Basic principles

Honing uses solid abrasives in the form of grinding blocks or sticks (*Figure 8.25*) and produces a highly characteristic surface which exhibits a cross-hatched pattern (*Figure 8.26*(*a*)). Honing is favoured for products requiring high accuracy and good surface finish and in bearing applications where the cross-hatched texture produced by the process improves oil-film retention.

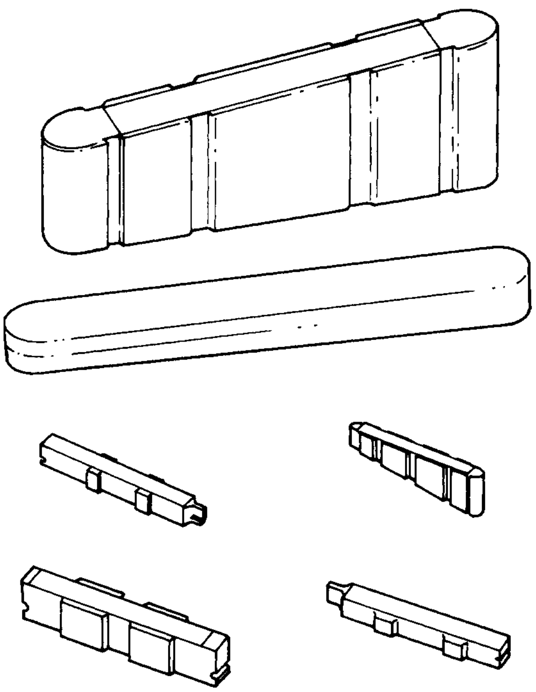


Figure 8.25 Examples of honing abrasives. (Courtesy of Jones and Shipman plc)

Typically, the honing process would involve machining a bore using specialist tooling of individual design which forces the grinding blocks or strips against the work piece. The tooling is subject to a complex motion consisting of both rotation and reciprocation. The cutting velocity V_c and the cutting angle θ can be calculated if the reciprocating velocity V_r and the rotational velocity V_t are known.

With reference to *Figure 8.26(b)*, the cutting angle θ is calculated as:

$$\tan(\tfrac{1}{2}\theta) = \frac{V_r}{V_t}$$

or

$$\theta = 2\tan^{-1}\left(\frac{V_r}{V_t}\right) \tag{8.16}$$

Also with reference to *Figure 8.26*(*b*) the cutting velocity is given by:

$$V_c = (V_r^2 + V_t^2)^{1/2} \tag{8.17}$$

or

$$V_c = V_t \cos(\tfrac{1}{2}\theta) \tag{8.18}$$

or

$$V_c = V_r \sin(\tfrac{1}{2}\theta) \tag{8.19}$$

Unlike most grinding processes, where the grits only cut for a brief period per revolution, in honing the cutting grits are in continuous contact with the work piece. However, the reciprocating motion of the process causes the individual grits to

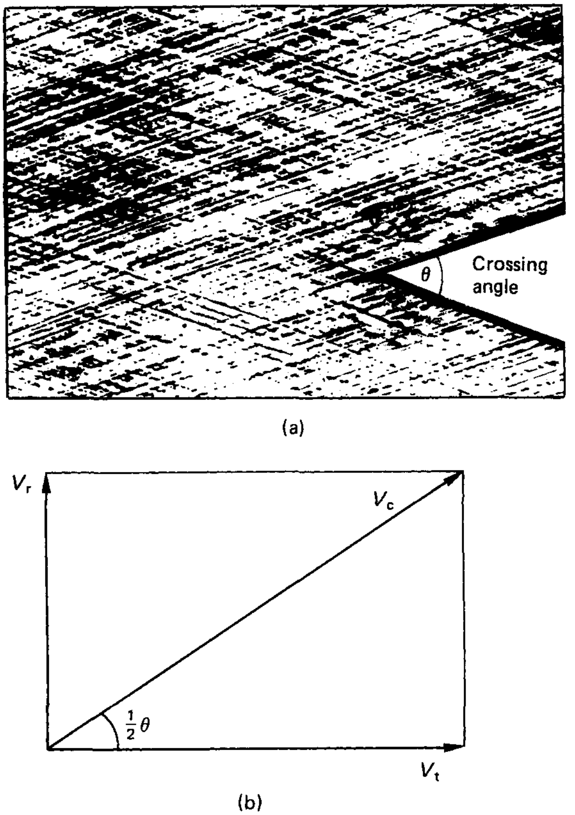


Figure 8.26 (a) Characteristic cross-hatched pattern produced by honing. (b) Diagram relating rotational velocity V_t and reciprocating velocity V_r to cutting velocity V_c

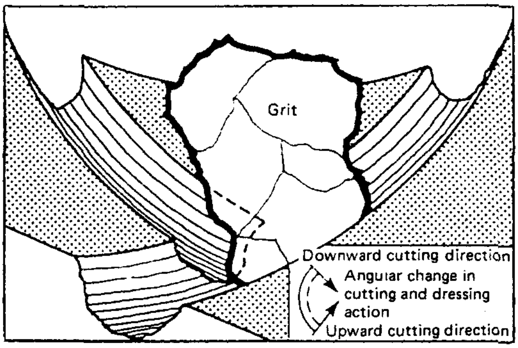


Figure 8.27 Effects of changes in direction on cutting facets. (Courtesy of Jones and Shipman plc)

describe a zig-zag path across the surface of the work piece, giving the characteristic cross-hatch pattern. These continual changes in direction alter the facets of the grit cutting at any one time, helping both self-sharpening and the cutting process (*Figure 8.27*). Honing is essentially a wear process in that a force is applied between the cutting stones and the work piece (similar to controlled force grinding) from which a cutting rate evolves. For any given set of cutting conditions (product), the life of the stone will depend on the relative area of stone to work piece which is defined as the stone face value α:

$$\alpha = \frac{A_s}{\pi D^2} \tag{8.20}$$

When optimum honing conditions have been established, it is useful to calculate the stone face value α for future reference.

8.14 Honing practice

8.14.1 Abrasives

Honing uses the complete range of abrasive materials including silicon carbide, aluminium oxide, diamond and cubic boron nitride (CBN). In the honing process, the abrasive is in continuous contact over a large area of the component. In order to achieve both self-sharpening and free cutting requires very careful selection of the abrasive type, grit size and bond hardness. The combination of grit and bond should satisfy the above requirements and also give an economic abrasive life.

The honing action constantly changes the point of shear load on each cutting grit, encouraging grit fracture as the cutting edges dull, thus new cutting facets or fresh grits are presented to the work surface. This self-sharpening action depends critically on selecting the appropriate abrasive grit and bond. If the bond is too weak the abrasive will wear rapidly giving a short abrasive life. Conversely, too strong a bond prevents self-sharpening leading to glazing, overheating and, as a result, a reduced rate of stock removal and poor geometry.

8.14.2 Tooling

Honing tools should be designed to ensure two essential features: that feed is applied equally to all the abrasive stones, and that abrading is uniform on all areas of the work after all high spots have been removed. Expansion is applied to each of the abrasives by a central cone, whilst the tool is rotated and reciprocated in the bore (*Figure 8.28*). As the abrasives contact the work piece, both cone and abrasives form a rigid cutting unit (*Figure 8.28*). To maintain stability, the abrasives are normally spaced evenly around the tool.

The tools which carry the abrading hones demand specialist attention at the design stage to ensure that consistent results are produced throughout their life. An individual approach to the design of quality tooling is essential as the characteristics of each product can vary considerably. The length of a bore affects: the length of abrasive; the diameter and the number of abrasives; and component section and the width and spacing of abrasives. All these factors need to be balanced correctly if optimum results and maximum efficiency are to be achieved.

Honamould tools are an example of the individual approach; this type of tooling extends from very small diameter to multibanked tools, which are used in conjunction with one of the following types of abrasive.

(1) Honamould: abrasive is moulded into a plastic material which supports and protects the abrasives from the harsh dressing action that is prevalent in extremely rough or interrupted bores.
(2) Diamond or CBN hones: the abrasive medium is sintered onto a metal strip machined to the required configuration.

Both these types of cutting medium are precisely manufactured to allow immediate use without the need for mounting

9

Fabrication

R D Cullum

E N Gregory

R Goss

Contents

9.1 Fasteners 9/3
- 9.1.1 Automatic insertion of fasteners 9/3
- 9.1.2 Joining by part punching 9/4
- 9.1.3 Threaded fasteners 9/4
- 9.1.4 Load sensing in bolts 9/5
- 9.1.5 Threadlocking 9/6
- 9.1.6 Threaded inserts and studs for plastics 9/7
- 9.1.7 Ultrasonic welding 9/7
- 9.1.8 Self-tapping screws 9/7
- 9.1.9 Stiff nuts 9/7
- 9.1.10 Washers 9/7
- 9.1.11 Spring steel fasteners 9/8
- 9.1.12 Plastics fasteners 9/8
- 9.1.13 Self-sealing fasteners 9/8
- 9.1.14 Rivets 9/8
- 9.1.15 List of suppliers of fasteners 9/10

9.2 Welding, soldering and brazing 9/11
- 9.2.1 Welding 9/11
- 9.2.2 Soldering and brazing 9/25
- 9.2.3 Productivity and welding economics 9/26

9.3 Adhesives 9/27
- 9.3.1 Introduction 9/27
- 9.3.2 Modern engineering adhesives by category 9/27
- 9.3.3 Designing with adhesives 9/28
- 9.3.4 Anaerobic adhesives 9/29
- 9.3.5 High-speed bonding 9/32
- 9.3.6 Toughened acrylics 9/32
- 9.3.7 Epoxies 9/33
- 9.3.8 Dispensing equipment 9/33
- 9.3.9 Health-and-safety considerations 9/33

9.1 Fasteners

Fasteners embrace a wide range of common and specialised components, but the emphasis for the designer of today is on fastening systems, largely because economy in production demands overall consideration of the assembly process. Fasteners inserted by hand are labour intensive and, therefore, make the assembly expensive.

Designers have to ask themselves which fastening system will serve the design purpose at the lowest overall cost. Does the design warrant fasteners and, if so, is it for assembly, maintenance, transport, etc? The reason for using a fastening device will lead to the choice of system, ranging from welding, through adhesives, bolts, rivets, etc., to highly specialised fastening devices. Despite the growing popularity of structural adhesives, mechanical fasteners are still widely used, but it must be emphasised again that the cost of inserting a fastener is likely to outweigh the cost of the individual fastener. Therefore, the designer must be convinced that a mechanical fastener is really the best solution for the job in hand. At this point it is convenient to enumerate questions such as:

(1) Is a fastener really necessary?
(2) Will the minimum number of fasteners be consistent with reliability and safety, as well as economic in production?
(3) Will the fastener specified perform its task efficiently?
(4) Will the fastener be simple to install and be capable of being placed by automatic methods?
(5) Will the fastener have to be removed during service for maintenance and repair, and will it be readily accessible with standard tools?
(6) If the fastener is to be removed during service, is it of a type that can be safely used again?
(7) Will the fastener material be compatible with that of the products?
(8) Bearing in mind the additional cost involved, will it be necessary to specify a special fastener instead of a standard one?

There are other questions that could be asked, and these depend on the type of assembly being considered, but the questions given above are basic to most designs.

The range of fastener systems currently available is bewilderingly wide, but it is the designer's responsibility to select from the right category, to assess performance from the supplier's data, and to consider in-place costs. Fasteners that fail in service are not only unreliable but uneconomical. The conditions under which the fastener operates must be known and the selected fastener must resist these conditions, which could well include extremes of temperatures, corrosive environments, vibration and impact loads.

9.1.1 Automatic insertion of fasteners

Depending on specific applications, most fasteners in use today can be installed automatically and, since some 80% of the total cost of fastening is on-assembly cost, a mere saving of 10% in assembly costs is more significant than a 40% saving in piece part costs.

Nevertheless, 80% of manufacturers assembling products with automated or robotic screwdriving equipment claim to have difficulties with their fasteners. This is the finding of a recent research survey carried out for European Industrial Services (EIS) who manufacture Nettlefolds screws. Faulty fasteners brought machinery to a halt and, in a significant number of instances, caused damage to machines or products.

For 23% of all users of automated equipment, the problem is a continuing one which production engineers appear prepared to live with, either because they are unaware that a solution exists, or because they believe that any solution must be too expensive. It is against this background that, in 1987, European Industrial Services introduced a fault-free product known as 'Nettlefolds Gold Seal' to the robot user market. However, it appears that just as many problems are being experienced by people with automatic assembly equipment. Now, EIS are offering the Gold Seal products for use with automated screwdriving equipment.

Virtually all the screw products used in robotic and automated screwdriving are hardened, mainly fastening metal to metal, but with a growing proportion joining plastics, or plastics to metal. The Gold Seal error scanned products, therefore, cater for these requirements with a variety of self-tapping screws for metal applications and Polymate, a specially designed EIS patented screw for thermoplastics.

The Pemserter is a high volume automated fastener insertion press from PEM International for inserting any of the range of bushes in the PEM range. The press has an intelligent main cylinder containing an optical encoder which transmits absolute ram position to a microprocessor based logic control system. Operating and safety windows coupled with an end-of-ram sensor allow the machine to perform at its maximum rate while providing operator and tooling safety.

During a set-up procedure, these safety windows are automatically determined. Any variation from the initial parameters such as the operator removing the anvil, or the presence of a foreign object such as a finger, will cause the ram immediately to stop and return. The company claim that their press achieves new standards in performance, safety and ease of operation.

Another example of a modern assembly tool is the fully automated twin headed Heatserter, which is capable of installing up to 45 fasteners per minute in plastics mouldings (see *Figure 9.1*). The new CNC machine from PSM International plc, can be preprogrammed in three axes (x, y and z) to cater for plastics mouldings of a wide range of shapes and sizes. Preheating circuitry is used to facilitate high-speed post-

Figure 9.1 PSM Automated Systems new CNC Heatserter can install up to 45 fasteners per minute and can be preprogrammed in three axes (x, y and z)

moulding fastener installation, and the equipment features light guarding to provide easy tool changing. The Heatserter can be interfaced with moulding presses to form a link in an automated production line and is capable of installing the full range of PSM proprietary fasteners for thermoplastics.

The problems of automated assembly are not just handling and orientation, but are often also those of quality and consistency. Design for automated assembly will, therefore, embrace both component design and the manufacturing process.

9.1.2 Joining by part punching

A very effective way of fastening sheets together is by part punching the top sheet into the bottom sheet. Material is partly cut from the top sheet and pushed through the lower sheet. Simultaneously, the lower sheet is formed so that an overlap between top and bottom sheets occurs. As the partial cut is connected to the upper sheet by webs, a join of considerable strength is produced.

The Trumpf TF300 power fastener has been developed to cut and form the sheets at the rate of 2 strokes/s and consists of a three part tool set comprising punch, stripper and die. Developed for static or portable use, an optional coordinate table ensures the accurate positioning of joints. The power tool is particularly useful where box shape enclosures are being assembled or where flat strips are to be joined together.

9.1.3 Threaded fasteners

Where threaded parts are concerned, designers have been bedevilled by the numerous thread forms that have been available through history and many are still with us today in spite of the UK's declared policy to recommend the adoption of the International Organisation for Standardisation (ISO) metric screw thread system. Existing thread systems include Whitworth, British Standard Fine (BSF), British Association (BA) as well as a range of pipe threads, gas threads, oil industry threads and fire service threads. However, there has been a part change typified by the use of ISO inch series of threads (Unified Coarse and Unified Fine). Nevertheless, wherever possible, new designs and updates should call for ISO metric thread systems.

The choice between Unified Coarse and Unified Fine threads depends on the complete understanding of the assembly in which screws and bolts are used. In general terms, the coarse thread (UNC) can fulfill most design requirements. One small practical advantage with the fine thread (UNF) is when they are used in conjunction with a slotted nut and split pin. To align the slot in the nut with the hole in the bolt, less torque is required with UNF for a given bolt tension.

Where bolts are used in tapped holes, a failure under tension loads should be by breakage at the core of the bolt thread rather than by stripping of the thread. Where bolts are in shear, designers should ensure that the loads are carried on the bolt shank rather than on the threaded portion. Another important aspect is the clearance of the bolt in the hole, particularly when several bolts appear in a row. The tolerance for dimensions must be carefully considered, otherwise undue strain will be put upon a single bolt rather than being equally distributed among them all.

With nuts and bolts there will always be a problem of making certain that in field replacements, the bolt removed is replaced by one of similar strength. Members of the British Industrial Fasteners Federation (BIFF) all produce bolts of high quality and standards and these can be readily identified from their head markings. Additional markings also indicate tensile strength and thread forms. *Figures 9.2* to *9.4* show the most commonly used head forms for screws, bolts and nuts.

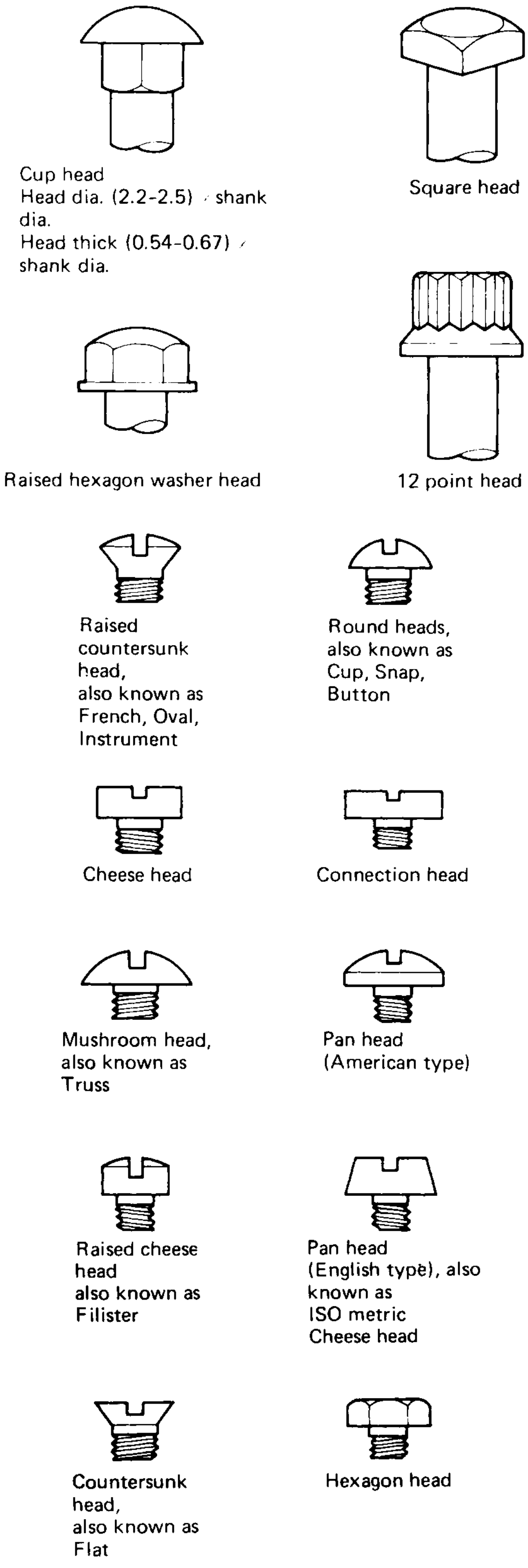


Figure 9.2 Head forms for bolts and screws

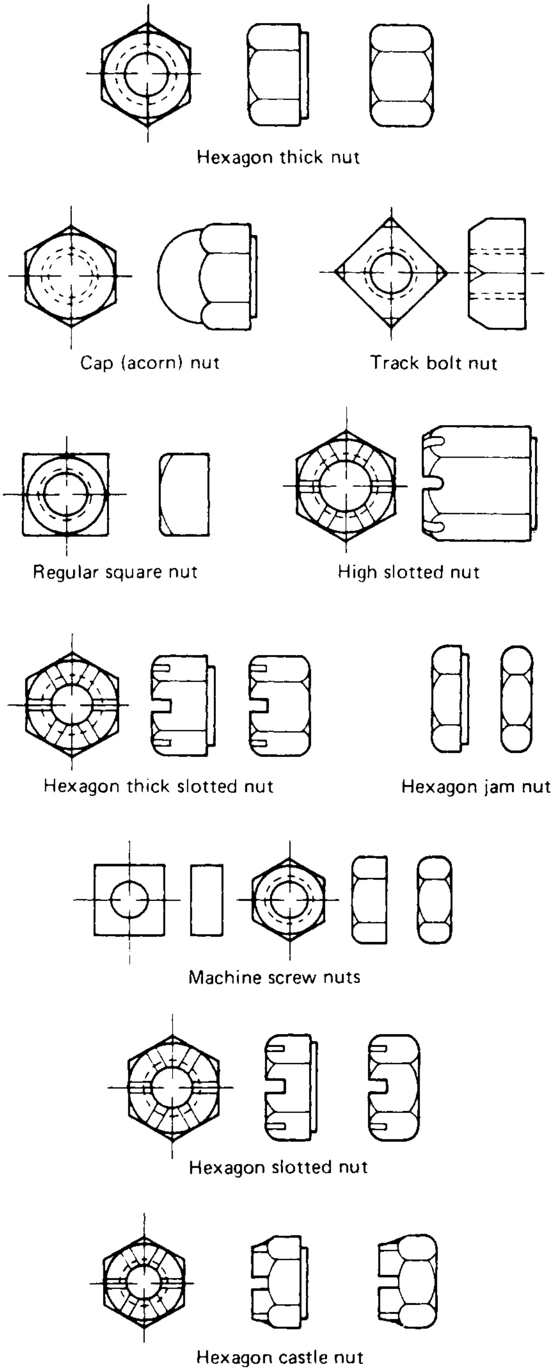


Figure 9.3 Free-spinning nuts

Threaded fasteners in general are used to provide a clamping force between two or more parts and where there is likely to be a need for dismantling at some time in the future, whether for access or for maintenance, etc. Clamping force can be a major problem, particularly when clamping soft materials which may shrink in service and cause loosening of the bolts. In cases like this, locking the nuts and bolts together may not always provide a satisfactory solution.

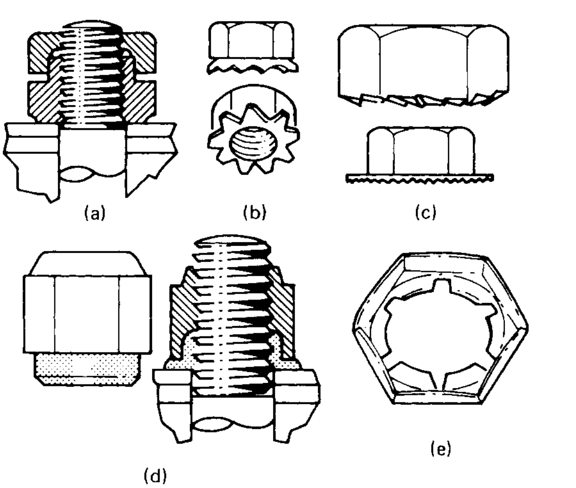


Figure 9.4 Free-spinning lock nuts. (a) The upper half of a two-piece nut presses the collar of the lower half against the bolt. (b) A captive tooth washer provides locking with a spring. (c) Ratchet teeth bite into a bearing surface. (d) A nylon insert flows around a bolt to lock and seal. (e) Arched prongs of a single thread lock the unit grip screw

9.1.4 Load sensing in bolts

Much research has been carried out on problems relating to bolt tension, and in many critical applications, ranging from car engine cylinder heads to sophisticated chemical and nuclear plants, it is necessary to apply the right amount of torque and this is usually done through the use of a torque spanner.

A clever idea to indicate bolt tension is to be found in the new fastener which is the brainchild of John Hirst who heads Exotech, a design and development company based at the University of Warwick's Science Park. West Midlands based T. W. Lench, one of the UK's leading nut and bolt manufacturers, is currently setting up a special manufacturing unit at its Warley factory to produce the new Hi-Bolt fastener, as it is called (see *Figure 9.5*).

The system involves a structural bolt with an in-built load sensing device. It is identical to a standard bolt apart from the actual sensor—a steel pin running through the shank which is gripped as the nut is tightened. Before tightening the pin is free to move. As the nut is tightened, the bolt shank contracts and grips the pin. When this happens, the maximum preload in the bolt has been reached.

The RotaBolt uses the same principle in a different way. The internal pin is secured at one end within the shank of the bolt. The free end holds a Rota washer and cap clear of the bolt head. As the bolt stretches under the tension imposed by tightening the nut, it pulls the Rota washer into contact with the bolt head and locks it. The gap between the Rota washer and bolt head has been set previously at the manufacturing stage so when the Rota washer locks, the predetermined tension in the bolt has been achieved.

To determine the actual bolt tension required for a particular application is far more complicated and, as previously mentioned, particularly so when dealing with soft materials such as glass fibre reinforced plastics (GRP). A research team at Imperial College undertook an extensive test programme to

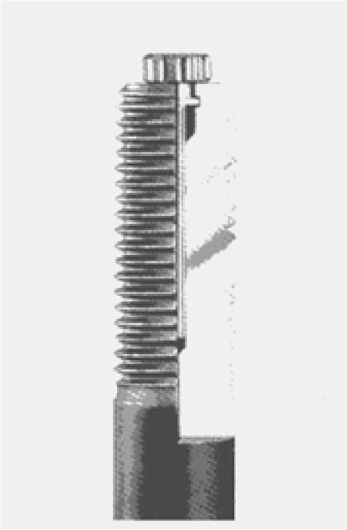


Figure 9.5 A close-up of the Hi-Bolt load sensing system from T. W. Lench

determine the static strength of bolted joints in several forms of GRP. The experimental results were used to produce design charts for both single and multibolt joints subjected to tensile/shear loading. While design methods established for structural joints in metals are applicable in a general fashion to GRP joints, the physical nature of the material does introduce problems not encountered with metals. The anisotropic stiffness and strength means that unexpected failure modes may be introduced, the low interlaminar shear and through thickness tensile strengths being a particular difficulty in this respect. Joint strengths were found to be strongly dependent on choice of reinforcing fabric.

Many technical devices have been used to prevent the unscrewing of nuts and bolts. A selection of ideas is shown in *Figures 9.8* and *9.9*. Each has its own advantages and disadvantages. Some can only be effectively used once, and once removed should be discarded. Serrations which effectively lock the bolt head or nut to a surface could initiate cracks in sensitive parts. In other cases, although the nut is locked, the bolt can turn. The most effective way of preventing relative rotation is by the use of liquid adhesives or anaerobic cements.

9.1.5 Threadlocking

The selection of a chemical threadlocking compound will depend on the following criteria.

(1) The ultimate shear strength of the fastener must not be exceeded, although this is usually only of importance for screw sizes of 5/16 in. diameter or less.
(2) The severity of the loosening tendencies.
(3) The size of the threads and, therefore, the viscosity of the compound to ensure thread filling.
(4) The method of application and requirements for testing and putting into service.
(5) The environmental requirements of temperature and chemical resistance.

The viscosity should be such that the compound can be easily applied, will not run off and will fill threads which have the maximum clearance. Preapplied dry materials are avail-

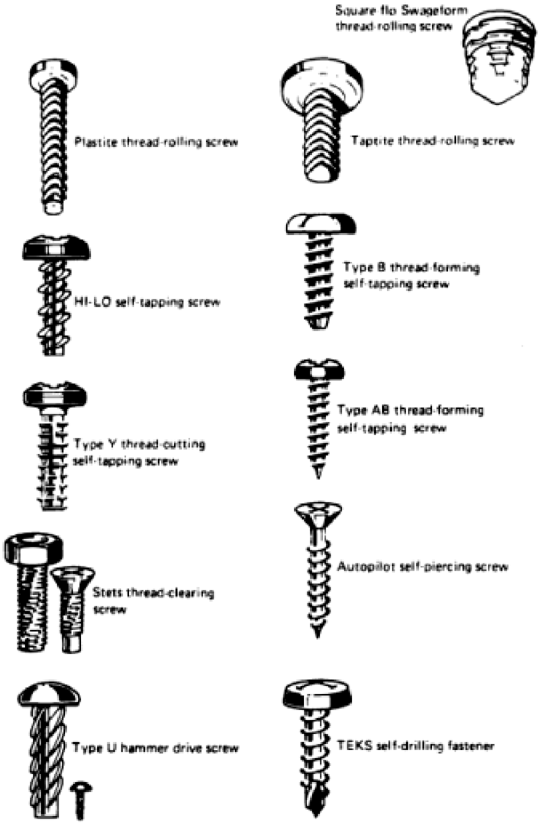


Figure 9.6 Self-tapping and thread-forming screws

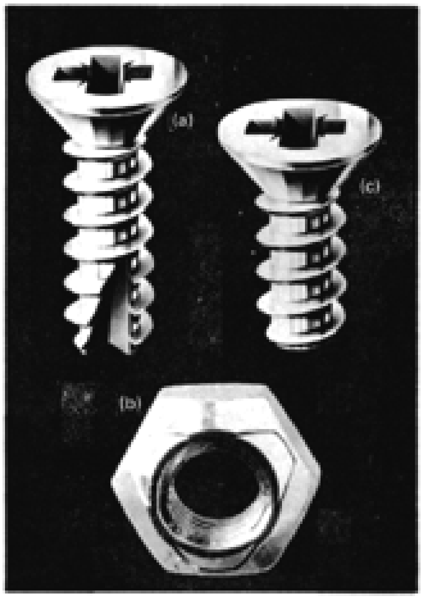


Figure 9.7 (a) Plasform K thread-cutting screw. (b) Reusable locking nut. (c) Plasform screw for plastics

able for application situations where liquids are not acceptable. They are easy to inspect and provide some reuse. The application is usually carried out by the bolt supplier and at least four formulations are available:

(1) low strength locking and sealing;
(2) medium strength plated fasteners;
(3) medium strength; and
(4) high strength.

PSM fasteners, for example, provide a comprehensive range of bolts and studs with the preapplied process for locking and sealing thread components. Their Scotch-Grip two-part epoxy system microencapsulates the resins while the hardener is freely carried in the coating substance. When the fastener is engaged with its mating thread, the capsules are crushed and the shearing action of the rotating fastener mixes the epoxy and hardener initiating the adhesive cure. An alternative but similar system uses microencapsulated anaerobic adhesives and sealants.

Fastening systems, whether they be adhesives, nuts and bolts, welding or any other method, are inherent in the design of the product or structure destined to be assembled as a one-off, by batch production or mass production. However, despite the popularity of adhesives, mechanical fasteners still have a prime role to play but, unless sufficient care is taken in their selection, the installation of such fasteners can lead to increased costs and possibly to a decrease in the required level of mechanical efficiency or in-service reliability.

To provide a comprehensive list of all of the various types of fastener available is impossible in this text, but it may be helpful to the reader to group the fastener types as follows:

(1) bolts, screws, studs, nuts and washers;
(2) self-tapping screws;
(3) locking and stiff nuts;
(4) blind bolts/screws;
(5) rivets including blind rivets;
(6) spring steel fasteners;
(7) plastics fasteners;
(8) quick release fasteners;
(9) self-sealing fasteners; and
(10) threaded inserts and studs.

Some of these are illustrated for general guidance, but each manufacturer produces its own wide range of products for similar applications. The reader would be wise, in the final analysis, to obtain specific data from these manufacturers, several of which are listed in Section 9.1.15.

9.1.6 Threaded inserts and studs for plastics

A method of securing a threaded bore in sheet is to use bushes designed to bite into the sheet and prevented from rotating either by being of hexagonal shape, or having a serrated shoulder or spigot. A large variety of these types of bush is available.

Whenever components must be fastened with screws to a plastics moulding, thread inserts which only require standard machine threads have distinct advantages:

(1) the assembly screw is completely reusable;
(2) unlike self-tapping or thread forming screws, which are applied directly into the plastics, there is no risk of thread stripping; and
(3) the assembly screw must often remain tight and maintain its clamping effect; threaded inserts allow this as they do not cause problems of stress relaxation in the material which is common when screws are driven directly into plastics.

The advantages of placing inserts after moulding instead of moulding-in are:

(1) reduced moulding time and, therefore, reduced unit costs;
(2) no expensive damage to tools due to misplaced inserts;
(3) no metal swarf from inserts to contaminate moulded surface; and
(4) availability of cost-effective assembly methods for post-moulded inserts.

Obviously, each job must be considered on its own merits, and post-moulded inserts tend to come into their own in long production runs. For small numbers, moulding-in can often be advantageous, in spite of the points made above.

9.1.7 Ultrasonic welding

Ultrasonics are widely used for joining plastics assemblies and for securing threaded inserts. The main limitation arises in joining parts made from different thermoplastics—a difference of only a few degrees in the softening temperature will cause one material to flow before the other. The pressure applied during ultrasonic welding is also a critical factor, and the most up-to-date ultrasonic welding equipment is microprocessor controlled to give the optimum pressure and heating conditions.

9.1.8 Self-tapping screws

These screws fall into two main categories: (a) true self-tapping screws which, like a tap, cut their own threads in a prepared hole; and (b) thread-forming screws which form their own threads by a rolling or swaging action (*Figures 9.6* and *9.7*). Typical of the latter are the proprietary types Taptite and Swageform. Taptite incorporates machine screw threads formed on a trilobular shank, and the Swageform has special lobes at 120° intervals along the tapered portion of the shank to impart a three-dimensional swaging action to form the threads.

Self-tapping screws are widely used in the assembly of sheet-metal components in which the thread-cutting action results in a good fit and is resistant to vibration and shock loads. Thread-forming screws are more applicable to thicker materials and blind holes. The self-drilling versions, for example Teks and Drill-kwick, are designed with a true drill point to produce the necessary pilot hole for the threads.

9.1.9 Stiff nuts

Stiff nuts, as opposed to free-spinning nuts, generally incorporate either a locking element to generate an elastic contact between the nut and bolt threads or, alternatively, a differential pitch or deformed threads to create the required friction between the mating threads. Examples are shown in *Figure 9.8*.

9.1.10 Washers

Washers are used to provide a seating for nuts and bolts in order to distribute the load over an area greater than that provided by the bolt head or nut. It also prevents damage to the surfaces being joined due to rotation of the nut or bolt. Washers can also seal, cover up oversize holes and act as a spring take-up between fastener and work piece. They also have the special task of preventing unwanted rotation of the nut or bolt. These include spring washers and serrated-tooth types as shown in *Figure 9.9*.

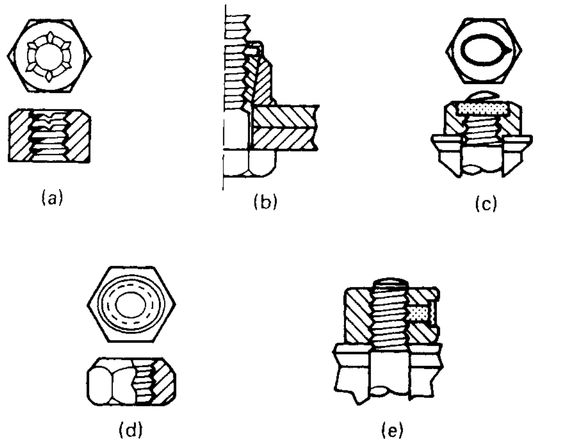


Figure 9.8 Stiff nuts. (a) The deformed thread has depressions in the face of the nut to distort a few threads. (b) The slotted section forms beams that deflect inward and grip bolt. (c) The threaded elliptical spring steel insert produces locking. (d) The non-metallic plug insert grips bolt threads. (e) The out-of-round threads cause wedging action

9.1.11 Spring steel fasteners

Although these include certain standard parts, for example, caged nuts, most spring steel fasteners are designed for specific applications where speed of assembly is important and the cost of the fasteners is low. They are mass produced from hardened and tempered steel strip and are used in non-critical applications in cars, domestic electrical appliances and in the sheet metal industries. Some of the various shapes which can be produced are shown in *Figure 9.10*.

9.1.12 Plastics fasteners

These products fall into two categories: (a) standard thread forms made from plastics such as nylon, and (b) specially configured fasteners for specific applications. Plastics fasteners can be self-coloured and are non-corrosive. Some examples are shown in *Figure 9.11*.

9.1.13 Self-sealing fasteners

Bolts, screws, washers and nuts can be made into self-sealing elements by the addition of a sealing element in applications where it is necessary to seal the fastener hole against the leakage of liquids or gases (see *Figure 9.12*). The choice of the fastener and the sealing material is dependent on the specific application.

9.1.14 Rivets

Rivets are low cost, permanent fasteners suitable for manual and automatic setting. They can be classified as solid, tubular, semi-tubular, bifurcated and blind, all of which are available in several different materials and head styles. Correct joint design and preparation together with the right fastener will ensure a strong trouble-free and cost-effective assembly. A range of available types is shown in *Figures 9.13* and *9.14*.

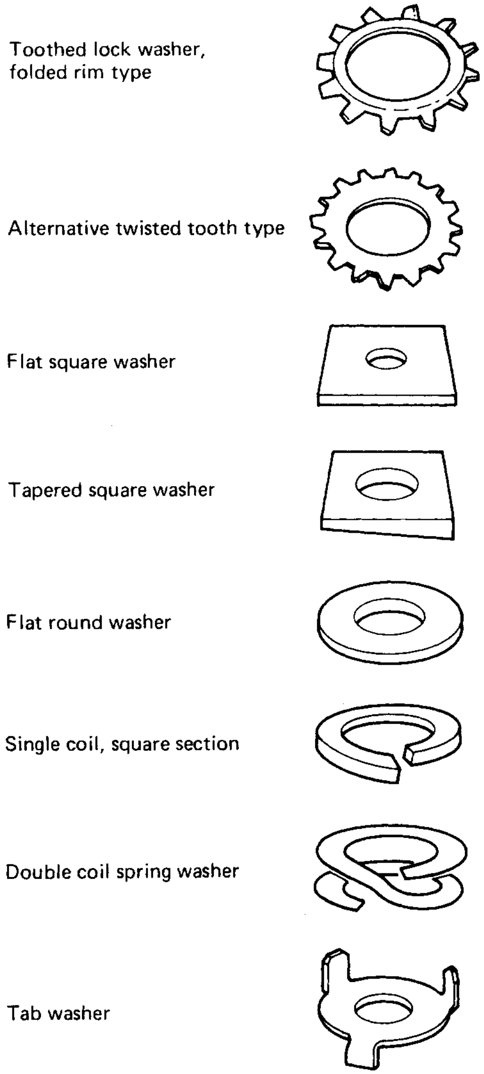


Figure 9.9 Washers

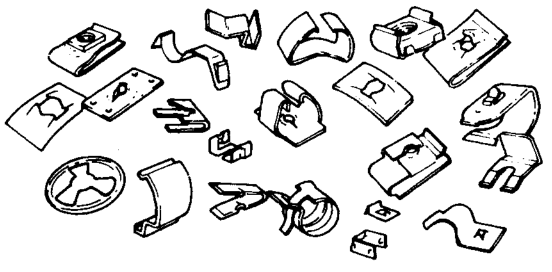


Figure 9.10 Selection of spring steel fasteners

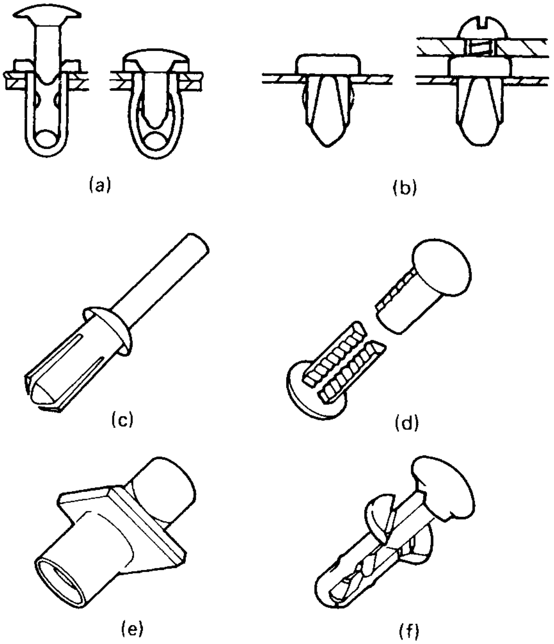


Figure 9.11 Various types of plastics fasteners. (a) Dart-type clip. (b) Nylon expansion nut. (c) Plastirivet. (d) Rachet rivet. (e) Plastics rivet. (f) Triploc rivet

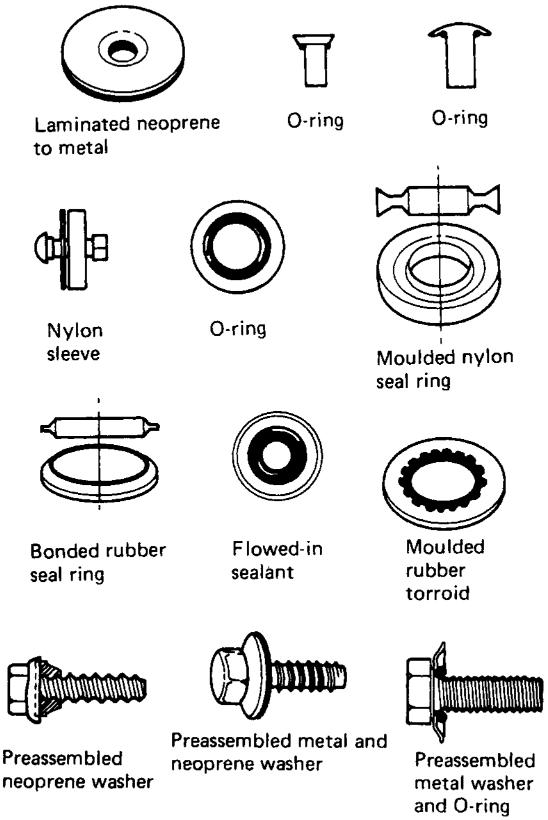


Figure 9.12 Self-sealing fasteners and washers

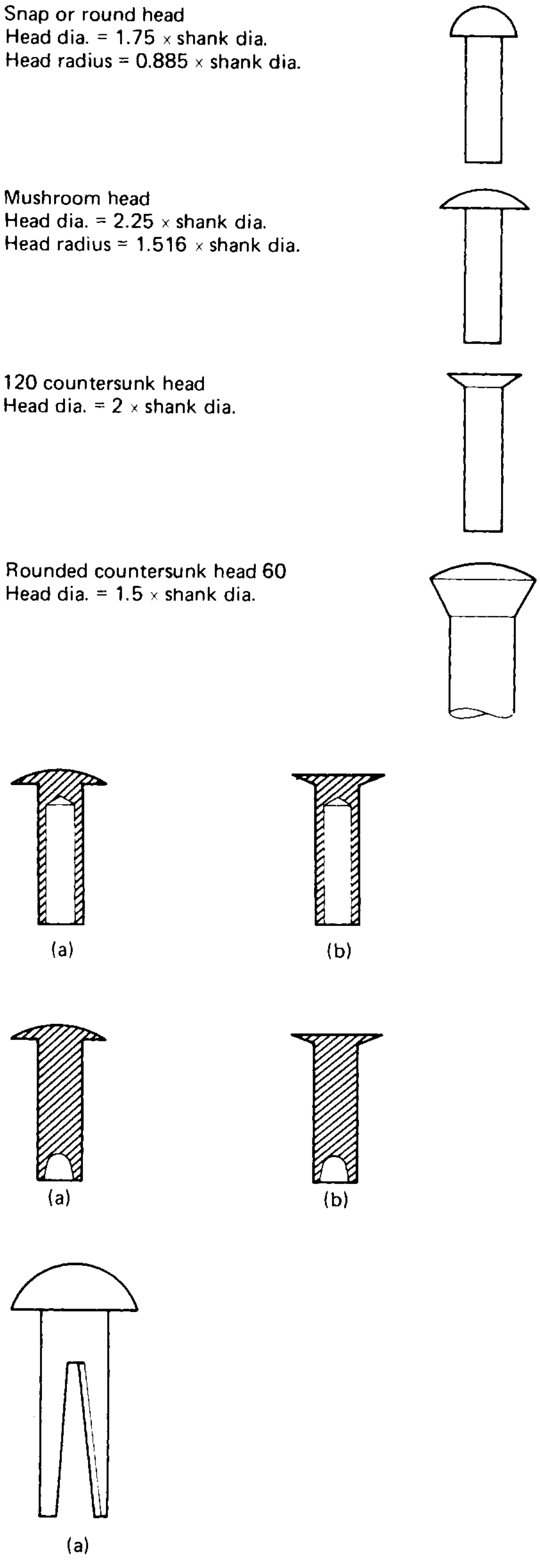


Figure 9.13 Selection of rivets

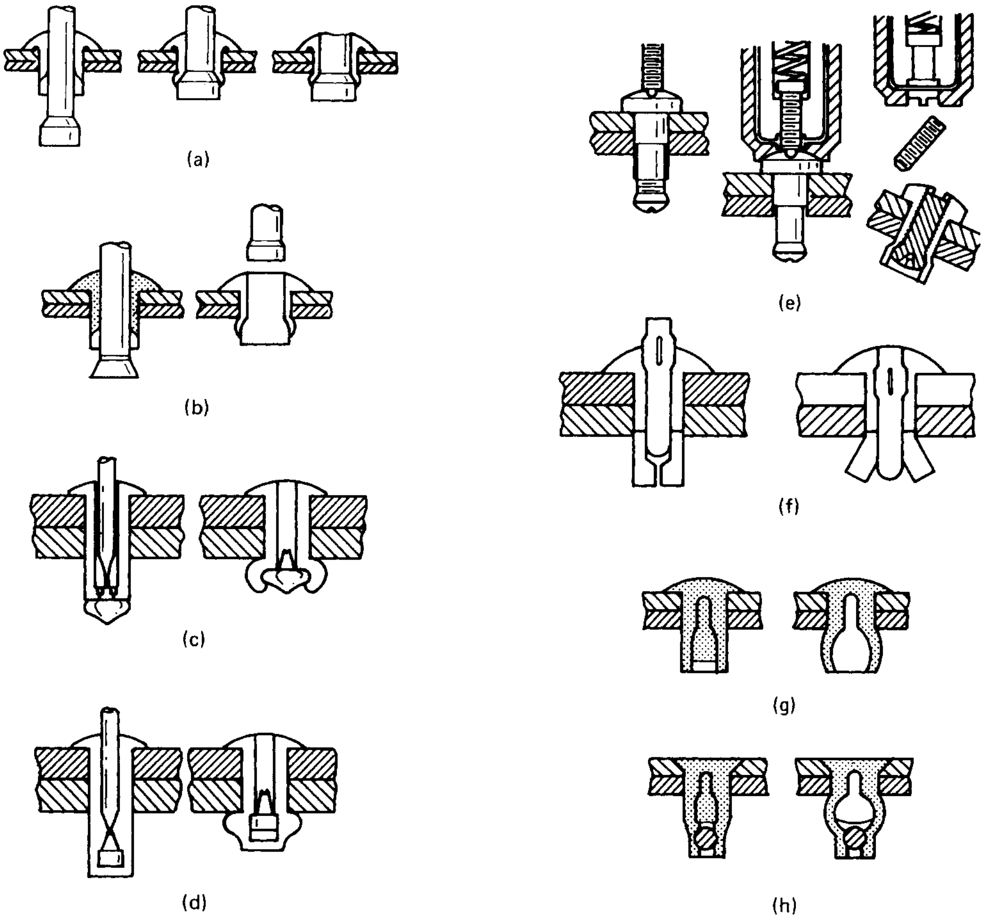


Figure 9.14 Blind rivets. (a) Break-stem mandrel. (b) Pull-through mandrel. (c) Break-stem mandrel (open end). (d) Break-stem mandrel (closed end). (e) Threaded mandrel. (f) Drive pin. (g) Chemical expanded (open end). (h) Chemical expanded (closed end)

9.1.15 List of suppliers of fasteners

Armstrong Screws and Fixings, 72 Great Barr Street, Birmingham B9 4BJ, UK
Bollhoff Fastenings Ltd, Midacre, The Willenhall Estate, Willenhall WV13 2JW, UK
British Industrial Fasteners Federation, Queens House, Queens Road, Coventry CV1 3EG, UK
Dzus Fasteners Europe Ltd, Farnham Trading Estate, Farnham, Surrey GU9 9PL, UK
European Industrial Services Ltd, Woden Road West, Kings Hill, Wednesbury, West Midlands WS10 7TT, UK
Evode Ltd, Como Road, Stafford ST16 3EH, UK
Fasteners Centre Ltd, The Hale House, Ghyll Industrial Estate, Heathfield, East Sussex TN21 8AW, UK
ISC Fasteners, 180 Rooker Road, Waltham Abbey, Essex EN9 1JJ, UK
Jetpress Sales Ltd, Nunn Brook Rise, Huthwaite, Notts NG17 2PP, UK
T W Lench Ltd, P.O. Box 31, Excelsior Works, Rowley Regis, Warley, West Midlands B65 8BZ, UK
Micro Plastics International Ltd, Unit 2 Healey Road, Coventry CV2 1SR, UK
PEM International, Kirk Sandall Industrial Estate, Sandall Lane, Doncaster, South Yorkshire, UK
PSM Fastener Systems Ltd, Longacres, Willenhall, West Midlands WV13 2JS, UK
SEAC Ltd, 46 Chesterfield Road, Leicester LE5 5LP, UK
Spiral Industries Ltd, Princewood Road, Corby, Northants NN17 2ET, UK
Southco Fasteners Ltd, Unit E, Gregorys Bank Industrial Estate, Gregorys Bank, Worcester, UK
TR Fastenings, Trifast House, Framfield Road, Uckfield, East Sussex TN22 5AR, UK
Tappex Thread Inserts Ltd, Masons Road, Stratford on Avon, Warks CV37 9NT, UK
Trumpf Machine Tools Ltd, Lyon Way, Hatfield Road, St Albans, Herts AL4 0LB, UK
Unbrako, Gunns Lane, West Bromwich, West Midlands B70 9HF, UK

9.2 Welding, soldering and brazing

9.2.1 Welding

Welding is used for joining metals so that the physical and mechanical properties of the parent metal are reproduced in the joint.

The integrity of a welded component, which has metallurgical continuity across the joint, is also characterised by properties such as pressure tightness, or heat and corrosion resistance. These properties have contributed to the rapid development, both technical and economic, in all fields including nuclear power, chemical engineering, bridge building, offshore engineering, shipbuilding, and the manufacture of automobiles, railway locomotives and rolling stock, aircraft engines, domestic appliances, and military hardware from small arms to main battle tanks.

Metallurgical continuity across a joint can cause problems compared with bolted or rivetted connections. For example, under certain conditions, cracks originating from fatigue or brittle failure can propagate across welded seams, whereas they are arrested by mechanical joints. Nevertheless, extensive research has been carried out over many years on the factors controlling fatigue and brittle failure, and guidelines are available in the literature and in various codes of practice which are effective in avoiding cracking problems during the service life of welded structures and components.

A weld is, in effect, a miniature casting, and some of the defects that can occur in castings may also be present in welds. The heat-affected zone adjacent to a weld is heated to high temperature during welding and cools rapidly as heat is conducted into the body of the component.

This area may be heated again to a lower temperature when more weld metal is deposited on the top of the first weld to build up the joint. The sequence of heating, cooling rapidly, and reheating to a lower temperature is similar to the heat treatment of steel by quenching and tempering.

Some welding processes involve substantial plastic deformation of metal adjacent to the interface between the parts being welded, which is similar in many respects to the forging process which is used to form metal parts.

Thus, welding encompasses the whole field of metallurgy, albeit on a small scale, and the application of metallurgical science has successfully solved the major welding problems such as cracking of weld metal or parent plate, porosity and slag inclusions in welds, as well as problems that may occur in service such as cracking due to fatigue, brittle fracture or creep, and wear or corrosion.

It will be readily understood that welding technology has developed into a discipline in its own right, and a short section can only refer to the main points. However, it is important for the engineer to know what methods are available for joining the materials used because the most economical manufacturing route in many cases depends on selection of the most appropriate welding procedure that will produce welded joints having the minimum level of quality fit for their intended purpose. This will require decisions to be taken by an experienced welding engineer who will generally have a professional qualification, but to understand the logic of the welding engineer's decisions a brief knowledge of welding technology is advantageous for any engineer. The purpose of this section is to provide basic information on the principal methods of welding, brazing and soldering and their fields of application. For more detailed information the reader should consult the references cited. It should be noted that in the early 1990s most of the British Standards referred to will be revised and given a European Standard (EN) number, although the British Standard number will not be changed.

9.2.1.1 Types of joint

The types of joint used and their associated weld types are described in detail in BS 499: 1983: Part 1.[1] The commonest joint types are butt, T, corner and lap.

Butt joints These are joints between parts that are generally in line. If two plates are placed in contact (a close square-butt joint) they can be welded with full penetration by one run of weld metal deposited by a manual welding process from each side provided that the plate thickness does not exceed approximately 8 mm (*Figure 9.15*). However, plate above 6 mm thickness is generally bevelled, and the V-edge preparation formed is filled by depositing a number of runs of weld metal. If high current mechanised welding processes are used, penetration of the weld may be at least double the above dimensions and for electron beam welding may be many times as high.

Lap joints These are commonly used for sheet metal up to about 3 mm thick in which one sheet is overlapped by another. This type of joint is used for soldering, brazing, resistance spot or seam welding, and for arc spot welding, plug welding, as well as for adhesive bonding. For material of 3 mm or thicker, even up to 10 mm lap joints are occasionally used and fillet welds are deposited at the plate edges by arc welding.

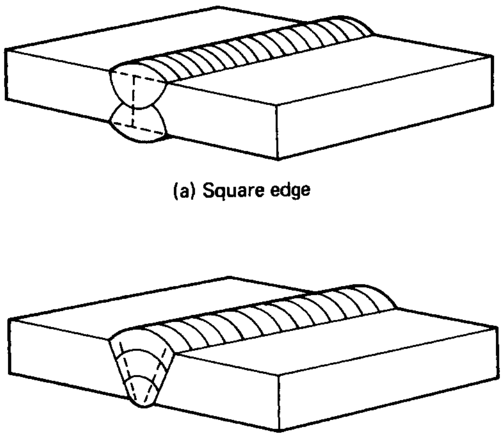
T joints and corner joints The parts may be joined by fillet welds or butt welds made by an arc welding process (*Figure 9.16*).

9.2.1.2 Welding processes

The various welding processes can be used to join the majority of metallic materials whether in cast or wrought form, in thicknesses from 1 mm or less up to 1 m or more.

A simple classification of welding processes is given in *Figure 9.17*. A complete classification and definitions of the processes are given in BS 499: 1983: Part 1.[1]

A description of the various welding processes is given below.



(a) Square edge

(b) Single bevel

Figure 9.15 Edge preparation for butt welds

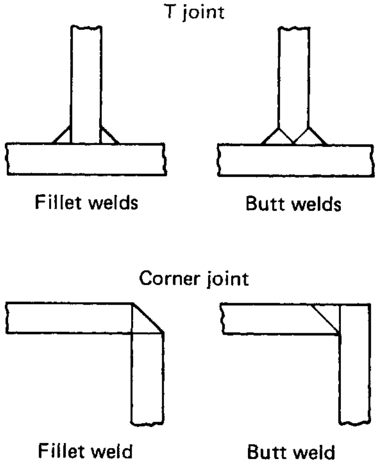


Figure 9.16 Examples of welded joints

9.2.1.3 Manual metal arc welding

Manual metal arc welding, referred to as 'shielded metal arc welding' in the USA, is the most widely used welding process and accounts for approximately 50% of all the welding carried out in the world today. With this process, welding is carried out with flux coated electrodes which are connected via an electrode holder and length of cable to one terminal of a welding power source, such as an a.c. transformer or a d.c. generator (*Figure 9.18*). The other terminal of the power source is connected to the work piece via the earth return or the ground cable, so that when the end of the electrode is placed in contact with the work piece, electric current flows through the circuit.

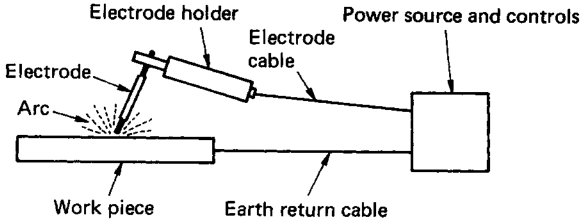


Figure 9.18 Welding circuit for manual metal arc welding

By withdrawing the tip of the electrode to about 3 mm from the work piece, an arc will be struck and current will continue to flow in the circuit and will pass through the arc which is electrically conductive.

If an arc is maintained between a rod type electrode and plates to be welded together, the tip of the rod becomes molten and so does a portion of the plates (the fusion zone). Gravity causes drops of molten metal to drip onto the plate and form a weld (*Figure 9.19*).

Apart from gravity, other forces caused by electromagnetic effects propel molten metal globules across the arc, and these forces always transfer metal from the rod to the plate whether

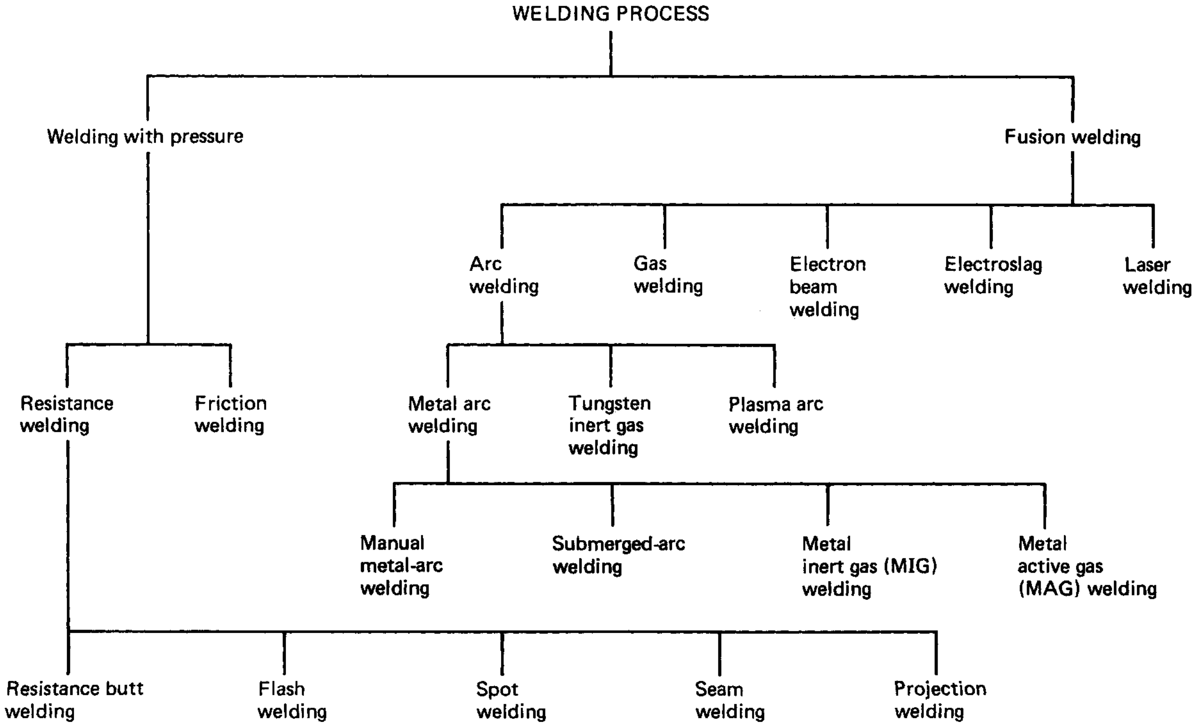


Figure 9.17 Classification of principal welding processes

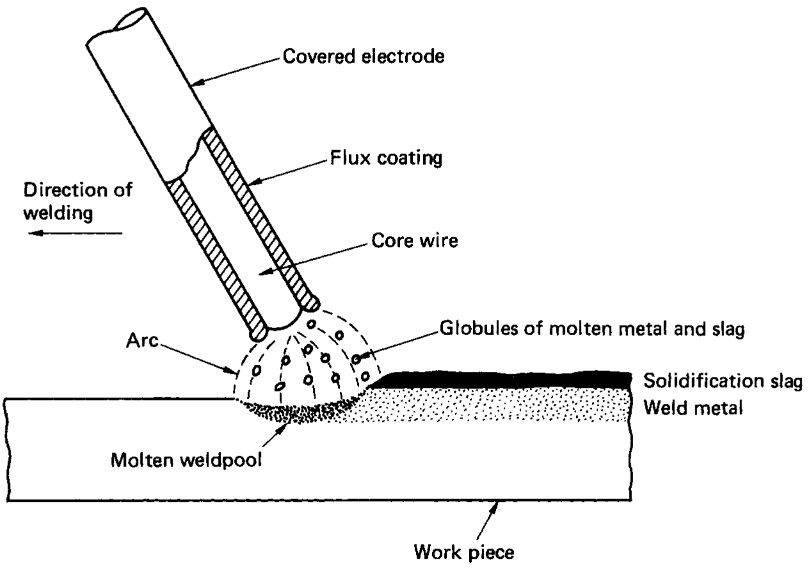


Figure 9.19 Manual metal arc welding with covered electrode

a.c. or d.c. is used and whether the polarity is electrode positive or negative. They will also transfer the metal against the force of gravity, so that vertical or overhead welding is possible.

Electrodes Electrodes have core wire diameters of 2.4–10 mm and are 300–450 mm in length. The deposition rate of weld metal, which governs the overall rate of welding, increases with increasing current and has a maximum value for each electrode length and diameter. Exceeding the maximum current causes overheating of the electrode core wire by resistance heating, which can damage the electrode coating. Welding currents vary from 60 A for the smallest electrodes, up to 450 A for the largest. The highest currents and deposition rates can only be used when welding downhand, i.e. in the flat position. Vertical and overhead welding can be used with electrodes having diameters up to 5 mm with maximum currents of approximately 170 A.

A bare wire can be used for welding, but the arc is mechanically unstable and the surface appearance of the weld is rough. The molten weld metal combines with nitrogen and oxygen from the air, resulting in poor mechanical properties.

The above problems can be overcome by coating wires with suitable fluxes and the principal purposes of this are:

(1) To facilitate the initiation or striking of an arc and to stabilise it so that it can be easily maintained.
(2) To provide a gas shield which protects the molten metal droplets from oxygen and nitrogen in the air as they are transferred through the arc.
(3) To provide a slag which protects the hot, solidifying metal from oxidation. The characteristics of the slag (e.g. melting point, surface tension, and viscosity) determine the shape of the weld bead and the suitability of the electrode for positional welding.
(4) To supply alloying elements to the weld metal; this means that an inexpensive rimming steel core wire can be used for many different weld metal compositions.

The constituents of the flux covering are mixed together in dry powder form and then binding agents are added. The weld paste is pressed into the form of slugs and loaded into machines which extrude the covering round electrode core wires as they pass at high speed through a die of appropriate size.

The electrodes are then dried as they pass through ovens and are stamped with identification marks before being packed.

The classification of flux coverings The development of flux coverings, consisting of mixtures of various minerals, has followed fairly well-defined lines with slight variations between different manufacturers and in different countries. Electrodes can be classified according to their coating types; for a full description BS 639: 1986[2] and the American standard AWS A5.1-81[3] should be consulted.

Steel electrode types are designated by letters in BS 639: 1986 and the chief characteristics of the different electrodes are as follows.

R (rutile)—Rutile coverings contain a high proportion of titanium dioxide in the form of the mineral rutile or ilmenite. The electrodes are easy to use but produce weld metal having high hydrogen contents which can cause cracking of the weld or parent metal heat-affected zone in heavily restrained joints.
RR (rutile, heavy coated)—The thick covering enables the electrodes to be used as contact electrodes which can be held in contact with the parent plate and dragged along the joint at high welding speed. Iron powder is often added to the coating to increase the deposition rate, and the RR type electrodes are not suitable for welding in the vertical and overhead position.
B (basic)—A basic covering usually has a high content of limestone (calcium carbonate) and fluorspar (calcium fluoride). Basic covered electrodes are often referred to as 'low hydrogen' because they were developed to produce weld metal having a low hydrogen content which reduces any tendency to hydrogen induced cracking. This covering decomposes to give a gas shield containing a large proportion of

carbon dioxide. These electrodes are used extensively because of their ability to weld medium and high tensile steels as well as high sulphur (free cutting) steels without solidification cracking of the weld metal and also because, by suitable drying treatment, the moisture content of the flux covering can be reduced so that the weld metal hydrogen content will be correspondingly low. This gives insurance against hydrogen-induced cracking of both the weld metal and the heat-affected zone (HAZ). Properly designed basic covered electrodes produce weld metal which has the highest fracture toughness properties, and they have the advantage over other types of electrodes in that high fracture toughness is maintained in all welding positions.

BB (basic, high efficiency)—These are similar to basic covered electrodes but have iron powder added to the coating so that the quantity of weld metal deposited is at least 130% of the weight of the core wire. The high deposition rate makes these electrodes unsuitable for welding in the vertical and overhead positions.

C (cellulosic)—This designation indicates a covering which has a high content of cellulosic material. These electrodes operate at a high arc voltage, which gives a deep penetrating arc and rapid burn-off. The covering forms a voluminous gas shield, consisting chiefly of carbon monoxide and hydrogen, and a small volume of slag which facilitates work involving changes in welding position such as pipe welding for which these electrodes are particularly suitable. They can also be used for the fast, vertical down welding of vertical seams in storage tanks up to about 12 mm thick. Because of the excellent penetration the root does not require gouging before making a sealing run on the reverse side. In pipe welding the close control of penetration is necessary because the deposition of a sealing run on the inside is generally impossible.

Power sources There are basically three types of power source:

(1) a.c. generators,
(2) d.c. rotary generators, and
(3) d.c. solid state.

The choice of power source is described by John and Ellis.[4]

A fourth type of power source of recent development is based on an invertor which is used to convert the mains frequency from 50 Hz to 5–25 kHz. Transformers for currents operating at these high frequencies are much lighter than those used in conventional a.c. generators. In the invertor-type power source, the a.c. mains input is rectified to give d.c. which is then fed to an invertor which converts it back to a high frequency a.c. The power is then reduced to the welding voltage by a lightweight transformer and it is rectified again to d.c. for welding.

Traditionally, in the UK transformers have been the most widely used type of power source for manual metal arc welding because of their relative cheapness, reliability and long life. The development of d.c. solid state power sources has eliminated the moving parts and maintenance costs associated with rotary generators, and has also reduced the capital costs.

Therefore, the use of d.c. welding is likely to increase in future, particularly in view of the fact that welding is easier because of the more stable arc and ability of d.c. to weld non-ferrous alloys and to produce the highest quality welds in stainless steel.

For use in repair work, for example in the garage trade, and for the 'do-it-yourself' market, a number of low-power welding sets are available which operate from the single phase 220/240 V supply. These power sources can be used with electrodes up to 3.25 mm diameter and both a.c. and d.c. units are available. For the home market, video instruction is available.[5]

Applications The chief reasons for the popularity of manual metal arc welding are its versatility, simplicity of operation, and the relatively low cost of equipment.

The process can be used with equal facility in a workshop or on site. To weld on a remote part of a structure, it is possible, within reason, to lengthen the cables from the power source and, when the limits of extended cables are reached, the power sources are readily transported by crane or motor vehicle, or manually on level sites, and movement of the power sources is facilitated by their simplicity and robustness.

The range of thicknesses welded varies from less than 2 mm in the fabrication of sheet metal ventilation ducting to 75 mm and above in the production of nuclear containment pressure vessels. These two examples are indicative of the wide range of quality standards that may be required, from general sheet-metal work up to the highest possible standard of radiographic soundness and mechanical properties.

Metals that are most commonly fabricated by manual metal arc welding are carbon and carbon manganese steels, low alloy steels and stainless steels of both the corrosion- and heat-resisting types. By selection of suitable electrodes described in various standards,[2, 6–9] the mechanical properties of the weld metal in respect of strength, ductility and toughness match those of the parent plate at ambient temperature and also at elevated or subzero temperatures, as required. References to American Welding Society (AWS) specifications are included because of the world-wide use of these in the oil and petrochemical industries.

Non-ferrous metals such as nickel, copper and aluminium and their alloys are welded much more extensively with the gas shielded processes, although nickel and nickel alloys are readily welded by the manual metal arc process and a wide range of electrodes is available.[10–12]

Some tin–bronze (copper–tin) and aluminium–bronze (copper–aluminium) electrodes are manufactured, but their chief use is for repair work, particularly of castings, e.g. marine propellers. These alloys can also be used for welding pure copper, because the high conductivity of copper has prevented the successful production of a copper electrode. Pure copper is generally used for either its high thermal or electrical conductivity and, therefore, the application of copper alloy electrodes is strictly limited to those circumstances where weld metal, having low thermal or electrical conductivity, is satisfactory. Some non-ferrous electrodes based on nickel, nickel–iron or nickel–copper alloys are used for welding cast irons.[14, 15]

A wide range of electrodes is available for hard surfacing components to increase their wear resistance under conditions of abrasion, impact, heat or corrosion or various combinations of these factors. Electrodes for hard surfacing are manufactured from core wires of mild steel, carbon and alloy steels, stainless and heat-resisting steels, nickel–chromium and cobalt–tungsten–chromium alloys, and are also made from steel tubes containing granules of refractory metal carbides such as tungsten and chromium carbides.[14, 15]

Gravity welding This is a semimechanised welding process which is used principally in shipyards for fillet welding stiffeners to horizontal plates. Covered electrodes of the contact type, typically 600 mm long, are supported by an electrode holder which slides down one arm of a tripod. The other end of the electrode is positioned in the corner of a T joint to be welded and when the current is switched on an arc is initiated and the electrode moves along the joint line as the electrode holder slides down the arm of the tripod under the force of

gravity. One man can operate three gravity welding units simultaneously, thus trebling the rate of manual welding.

Open arc automatic welding Although no longer used, it is appropriate to mention a mechanised welding process which was used extensively in the 1950s up to the 1970s in shipbuilding and bridge building. The engineer may find the process referred to in periodic inspection reports of these types of welded structures.

Mechanised welding with covered electrodes was carried out with a continuous coiled electrode having a core wire of 4–8 mm diameter. The core wire was wrapped helically with two thin wires (about 1 mm diameter) which anchored the extruded flux in place and also acted as a means of conducting electrical current from the jaws of the welding head to the core wire.

Automatic open arc welding with continuous covered electrodes has now been superseded by the submerged arc process.

9.2.1.4 *Submerged arc welding*

This is the most widely used mechanised welding process (*Figure 9.20*). A base wire (1.6–6.3 mm diameter, but usually 3.25 or 4 mm) is fed from a coil and an arc is maintained between the end of the wire and the parent metal. As the electrode wire is melted, it is fed into the arc by a servocontrolled motor which matches the wire feed rate to the burn-off rate so that a constant arc length is maintained.

The region of the joint is covered with a layer of granular flux approximately 25 mm thick, fed from a hopper mounted above the welding head. The arc operates beneath this layer of flux (hence the name 'submerged arc'). Some of the flux melts to provide a protective blanket over the weld pool and the unmelted flux is collected and reused. The electrode wire, welding head, wire drive assembly and flux hopper are mounted on a traverse system which moves along the work piece as the weld metal is deposited.

The traverse system may consist of a carriage mounted on a boom or it may be a motorised tractor either on rails or running freely with manual adjustment to follow the weld seam. Alternatively, the welding head can remain stationary while the work piece is moved. This method is used for welding the circumferential seams of a pressure vessel while it is rotated under the welding head.

Electrode wires The electrode for submerged arc welding is a bar wire in coil form usually copper coated. Two types are available: (a) solid wire, and (b) tubular wire. The solid wire is widely used for general fabrication of mild and low alloy steels, stainless steels and non-ferrous metals. For welding mild steel and low alloy steels, it is either a low carbon ultra-low silicon steel or a silicon-killed steel with manganese addition and sometimes low alloy additions, the selection of either type depending upon the type of flux to be used with it, i.e. a flux with manganese or manganese and alloy additions or a neutral flux, respectively. The tubular wire (made by forming narrow strip into a tube) carries alloy powders which permit the economical production of a wider range of weld compositions than is possible by using the solid wire type. Tubular wires are widely used for hardfacing.

Wire compositions for welding carbon steel and medium tensile steel are listed in BS 4165: 1984.[16]

With coated manual electrodes, wire and coating are one unit so that such electrodes can be classified according to the type of coating and its effect on weld mechanical properties. In submerged are welding, any wire may be used with a number of different fluxes with substantially different results in respect of weld quality and mechanical properties.

Consequently, BS 4165: 1984 grades wire flux combinations according to the tensile and impact strengths obtained in the weld metal.

A number of tubular wires are available, particularly for surfacing and hardfacing. These contain alloy powders which produce weld metals consisting of low alloy steels, martensitic and austenitic stainless steels, chromium and tungsten carbides, and various cobalt and nickel based heat- and corrosion-resistant alloys. Some corrosion-resistant alloys, including stainless steel are available in the form of coiled strips (100–150 mm wide, 0.5 mm thick) for high deposition

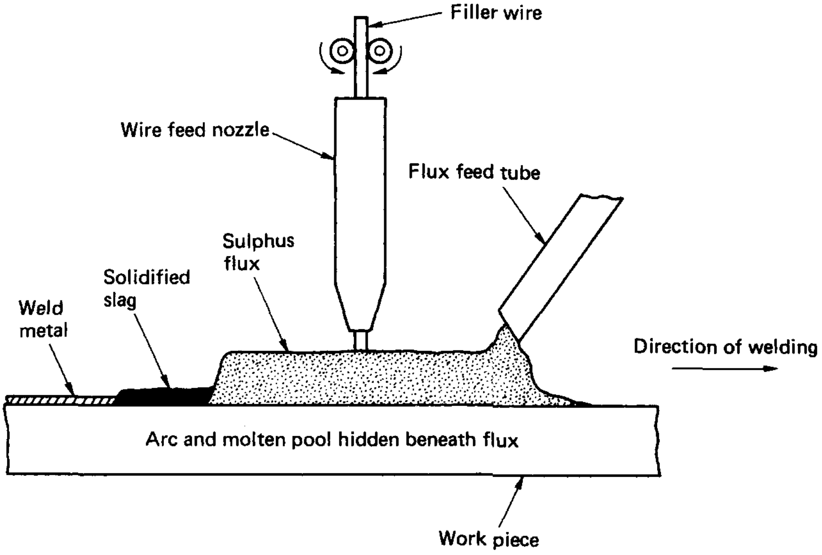


Figure 9.20 Submerged arc welding

rate surfacing by a submerged arc welding process known as 'strip cladding'.

Fluxes Two main types of flux are available, fused and agglomerated. Fused fluxes are manufactured by fusing together a mixture of finely ground minerals, followed by solidifying, crushing and sieving the particles to the required grain size. Fused fluxes do not deteriorate during transportation and storage and do not absorb moisture. Agglomerated fluxes are manufactured by mixing finely ground raw materials with bonding agents such as sodium or potassium silicates followed by baking to remove moisture. This type of flux is sensitive to moisture absorption and may require drying before use. Agglomerated fluxes are more prone to mechanical damage which can cause segregation of some of the constituents.

Fluxes are classified as acid, neutral, or basic, the latter being subdivided into semi-basic or highly basic. The main characteristics of the fluxes are as follows:

Acid fluxes—These have a high content of oxides such as silica or alumina. They are suitable for high welding currents and fast travel speeds. They are resistant to porosity when welding rusty plate, and have low notch toughness. They are not suitable for multipass welding of thick material.
Neutral fluxes—These have a high content of calcium silicate or alumina rutile. They are suitable for fairly high welding currents and travel speeds and also for multipass welding.
Basic fluxes—These have a high content of chemically basic compounds such as calcium oxide, magnesium oxide and calcium fluoride. They give the highest weld metal quality in terms of radiographic soundness and impact strength. Lower welding currents and travel speeds are suitable for multipass welding of thick sections. For further information on fluxes see references 16 and 17.

Power sources Either d.c. or a.c. may be used. D.c. may be supplied either by a motor generator or by a rectifier, either of which can have flat or sloping current voltage characteristics. These are referred to as 'constant-voltage' or 'constant-current' types. Generators will deliver up to about 650 A continuous output, and rectifiers deliver up to about 1200 A at 100% duty cycle. A.c. is supplied by welding transformers which are designed to give a drooping characteristic. Transformer output may range up to 2000 A.

Power sources are of rugged construction and are designed for 100% duty cycle. With the constant-current type source the arc voltage determines the wire feed rate which varies to maintain a constant arc length. The constant-voltage type power source produces a self-adjusting arc in which an increase in arc length or arc voltage causes a decrease in current and burn-off rate so that the original arc length is rapidly obtained. A decrease in arc length increases the current and burn-off rate and this self-adjusting effect occurs with a constant wire feed rate.

Application As the process operates with a continuous coil of electrode wire, butt welds in the flat position requiring multiple runs to fill the joint can be made with minimum stops and starts. Thus, circumferential joints in cylindrical bodies such as pressure vessels, pipes, etc., can be made with one stop and start per revolution of the work piece, this stop being necessary to reset the position of the welding head. Consequently, the possibility of stop and start defects is minimised; a most important consideration when reliability in costing is required. Although most widely applied to welding of joints in mild steel, low alloy high tensile steel, creep resisting steels, and to a lesser extent stainless steels, it is also widely used for building-up work, either for reclamation or replacement of defective parent metal or for hard surfacing.

Submerged arc welding is suitable for welding material of 5–300 mm thickness, and even thicker, but plates less than about 10 mm thick are generally welded by the gas shielded or flux cored arc welding process. A semi-automatic variant of the process is available in which the welder manually manipulates a welding gun on which is mounted a small hopper containing the flux. Electrode wire is fed to the gun from a coil by a wire feed unit. This process, which is only used to a limited extent, is sometimes referred to as 'squirt' welding.

Other variations of the submerged arc welding process are mainly concerned with increasing deposition rates and, therefore, welding speed and productivity. These include the following:

(1) Increasing the electrode extension or stick out by up to 150 mm by using an insulated guide tube. The resistance of the wire increases the burn-off rate by the I^2R heating effect.[18]
(2) Addition of iron powder to the joint which increases the weld volume.
(3) The use of multiple wire techniques in which two or more wires are used with two or three separate power sources.[18]
(4) Narrow gap welding of plates more than 100 mm thick in which a parallel gap of 14–20 mm between the square edges of the plates is used instead of a V or U groove. The saving in quantity of weld metal used is considerable, with a consequent increase in productivity.[19]

9.2.1.5 *Electroslag welding*

Electroslag welding is an automatic process for welding material 18 mm or thicker in the vertical position. Square edge plate preparation is used and joints are limited to butts and T butts. Heat for fusion is obtained by the current in the consumable electrode wire which passes through the molten slag formed by melting of a flux formulated to have high electrical resistance in the molten state.

The joint is set up with a wide gap (approximately 25–36 mm, depending on plate thickness), the very large weld pool being contained in the joint by water cooled copper shoes. The joint is filled in one pass. One, two or three wires are fed into the joint with or without a reciprocating motion to ensure uniform heat generation, the plate thickness determining the number of wires. One welding head with three wires can weld plate 450 mm thick. The copper shoes rise up the joint to prevent spilling of metal and slag. These shoes form part of the welding head.

A variation of the electroslag process is known as 'consumable guide welding'. Here the welding head remains stationary and feeds one or more wires down a tube which melts into the pool. The equipment is substantially cheaper than the conventional slag welding machines.

The consumable guide process, while theoretically suitable for very thick sections, is more usually applied to metal of thicknesses up to, say, 50 mm, because it is more manageable in these thicknesses. The equipment needed is much simpler than for the conventional process, consisting of a constant potential generator, rectifier or transformer with a flat characteristic and a wire feed unit. These are also the essential ingredients of submerged arc equipments and, by slight modification, they can be adapted for consumable guide welding.

The ultra-slow cooling rate of an electroslag or consumable guide weld minimises hydrogen cracking suceptibility in the parent steel and weld metal, but produces a large grain size weld. This tends to give a comparatively poor notch impact

strength as determined by conventional tests. Consequently, for pressure vessel application, current codes require normalising after welding.

Fluxes Electroslag welding fluxes produce complex silicate slags containing SiO_2, MnO, CaO, MgO and Al_2O_3. Calcium fluoride is added to increase electrical conductivity and lower slag viscosity. Slags based on CaF_2–CaO have a strong desulphurising action which assists the welding of steels having carbon contents greater than 0.25%, without solidification cracking in the weld metal. Fluxes must be kept dry.

Economics of the process On heavy steel plate, for pressure vessels, boiler drums, etc., the actual welding speed (rate of filling the joint) is about twice as fast in 40 mm plate, four times as fast in 90 mm plate, and eight times as fast in 150 mm plate as compared with multirun submerged arc welding. The welding speed is 1–1.7 m h^{-1}. Plate-edge preparation by bevelling is also avoided. On this evidence the process would seem highly attractive economically speaking. However, 'setting up' the machine and selecting the correct welding parameters is a matter of experience or tests. Therefore the 'setting-up' time must also be considered and the cost of determining the correct procedure included in the cost estimate.

Thus, the process gives full economic benefit on repetition work where set parameters can be used based on experience. An example is the wide use of the process for the longitudinal seams of boiler drums in 125–150 mm thick steel. On one-off applications involving plate thicker than 150 mm, the process is economical on each joint and, providing sufficient work of suitable type is available, the high capital cost is recovered within a reasonable time.

Mechanical properties The weld metal mechanical properties are determined to a considerable degree by the composition and cleanliness of the parent steel. However, proper selection of wire and flux and suitable parent metal confer strength and ductility equal to or better than the parent metal.

In the as-welded condition, degradation of notched impact properties in the heat-affected zone results in values lower than the weld metal. Where notched impact requirements must be met, it is necessary to normalise the completed joints. In general, electroslag welding is an acceptable and economical method of welding thick steel. It finds application for ships' hulls, boiler drums, press frames, nuclear reactors, turbine shafts, rolling mill housings and similar heavy fabrications.

9.2.1.6 Gas shielded metal arc welding

This general term covers a group of welding processes in which no added flux is used. The molten weld pool is protected by a gas shield which is delivered to the welding gun through a flexible tube at a controlled rate, either from a gas bottle or from a bulk supply. The shielding gas may be inert (e.g. argon or helium or mixtures of these gases) or it may be active (e.g. carbon dioxide (CO_2) or mixtures of CO_2 with other gases such as argon). Sometimes small additions of oxygen or hydrogen are included in the shielding gas.

The wide variety of shielding gases, which may be completely inert (or non-reactive) or active (i.e. slightly oxidising or reducing), has led to the use of the terms 'metal inert gas' (MIG) and 'metal active gas' (MAG) to describe the principal gas shielded metal arc welding processes.

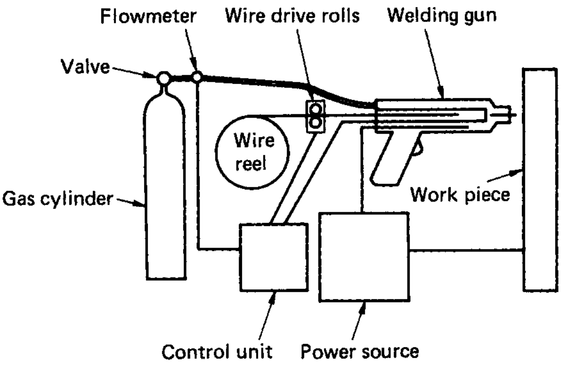


Figure 9.21 Metal inert gas (MIG) welding

9.2.1.7 MIG and MAG welding

In MIG or MAG welding, referred to as 'gas metal arc welding' (GMAW) in the USA (*Figure 9.21*), a small diameter (0.6–1.6 mm) wire is fed from a coil by a wire feed unit which contains an electric motor, gearbox and grooved drive rolls. The wire is fed to a welding gun which has a trigger which operates the wire feed drive, the current and the flow of shielding gas.

An arc is struck when the wire contacts the work and the arc length depends on the voltage which is preset by adjustment of a knob on the power source. Welding current is picked up by the wire from a copper contact tube through which the wire passes. The distance between the contact tube where current enters the wire and the end of the wire is usually a maximum of 25 mm, compared with 300–450 mm for a covered electrode. Therefore, overheating is not a problem, particularly as there is no flux coating. Higher currents can be used than those normally used for manual metal arc welding, e.g. 120–450 A for 1.6 mm diameter wire.

Therefore, deposition rates are generally higher than for manual metal arc welding. Another advantage of MIG/MAG welding is that the arc is automatically maintained at a length which depends on the arc voltage; thus the welder only has to move the welding gun along the joint line holding the nozzle of the gun at approximately the same distance from the joint. This is because the arc is self-adjusting because the voltage current characteristic of a MIG/MAG power source is flat or only slightly drooping.

If the welding gun is moved away from the joint the arc length and the arc voltage increase slightly. With a flat or slightly drooping characteristic, a small increase in voltage will cause a large decrease in welding current and the wire will burn off at a lower rate. The original arc length will be rapidly attained at which voltage the burn-off rate will once again match the wire feed speed.

Similar self-adjustment in the opposite sense will occur when the welding gun is moved towards the work piece.

Filler wires The commonest filler wires are 1.0, 1.2 and 1.6 mm in diameter, with 0.6 and 0.8 mm diameter wires less frequently used. Because of the wide range of current that can be used with each wire, it is only necessary to stock one or two diameters. This is in contrast with manual metal arc welding where a number of different diameters of electrodes and possibly two or more coating types may be required just for welding a single type of material such as carbon manganese steel.

However, a disadvantage of solid filler wires is the limited range of compositions available, because it would be too expensive for a steelmaker to produce small quantities of low alloy or stainless steel wires. Small batches of covered electrodes can readily be produced by introducing alloys in powder form through the coating. This situation is reflected by the number of electrodes and filler wires for welding low alloy and stainless steels listed in British Standards, which are as follows:

	Low alloy steels	*Stainless steels*
Covered electrodes	30	39
Solid wires	6	16

Any deoxidising elements such as silicon or aluminium required to refine or degas the weld pool are contained in the solid wire and compositions of wires available are listed in BS 2901: Parts 1–5[20] for ferritic steels, austenitic stainless steels, copper and copper alloys, aluminium and aluminium alloys and magnesium alloys, and nickel and nickel alloys.

Modes of metal transfer In MIG or MAG welding the operating conditions in terms of current and voltage determine the type of metal transfer which must be suitable for the application. There are four modes of metal transfer.

(1) *Short circuiting (dip transfer)*—This occurs when a low voltage and current are used which causes metal to be transferred from the end of the wire to the work piece by frequent short circuiting of the wire to the weld pool. This technique produces low heat input and a small controllable weld pool essential for welding steel sheet in all positions and thicker steel sections in the vertical and overhead positions. A disadvantage of the process is the production of spatter in the form of globules of metal expelled from the weld pool when each short circuit is broken. Spatter particles which adhere to the work piece can be reduced by fine tuning of the inductance of the power source.

(2) *Globular transfer (semi-shorting)*—This occurs when somewhat higher currents and voltages are used than for dip transfer welding of steel, but metal transfer still occurs by short circuiting of the filler wire to the weld pool. Because of the large droplet size and the larger weld pool, this mode of welding is not suitable for vertical or overhead welding. The production of spatter still occurs.

(3) *Spray transfer*—Free flight of metal droplets occurs with no short circuiting when the current and voltage are sufficiently high. This gives maximum deposition rates and deep penetration welding suitable for flat position welds in medium and heavy steel plate and for horizontal-vertical fillet welding (e.g. between a vertical and a horizontal plate). Spray transfer is used for welding aluminium and aluminium alloys in all positions because the spray transfer of droplets occurs at much lower welding currents than with other metals. Therefore, small weld beads can be deposited which solidify rapidly and enable welding to be carried out in the vertical and overhead positions.

(4) *Pulsed transfer*—This was developed to produce spray transfer at all current levels so that welding of all metal thicknesses in all welding positions could be carried out without the formation of spatter. In pulsed transfer the welding current is switched from a high pulse current to a low background current at a typical frequency of 50 Hz. The background current is sufficient to sustain the arc but it is insufficient for metal transfer. The pulse current is set above the critical level to produce sufficient electromagnetic force with each pulse to transfer one metal droplet from the tip of the wire. With the first pulsed arc power supplies the pulse frequency had to be a multiple of mains frequency and setting up welding conditions was difficult to the extent that it hindered the use of the process in industry. The average current, which depended on the background current, the pulse current and the frequency had to produce a usable burn-off rate at a constant arc length. The process was also sensitive to electrode stick out (the electrode extension beyond the contact tube) which could disrupt the balance between pulse energy and metal transfer, causing arc extension and spatter. The full advantages of the pulsed MIG process including stable low mean current operation, particularly when welding aluminium alloys or stainless steel and positional welding capabilities of all metal thicknesses were made readily available with the development of transistorised power sources referred to in the next section.

Power sources MIG and MAG welding are always carried out with d.c. and the principal types of power source are transformer rectifiers with constant potential or controlled slope characteristics and motor generators which are used for site work, e.g. welding pipelines.

Invertor-type power sources, described in Section 9.2.1.3, are also used for MIG and MAG welding. The main advantage of invertors is the considerable decrease in size and weight compared with conventional transformer rectifiers.

Electronic power control has had a considerable and beneficial influence on MIG/MAG welding enabling the process parameters to be preprogrammed which eliminates the complicated setting up operation and enables 'one-knob' control to be achieved.

Programmed control for both dip and spray transfer was originally developed in the late 1960s. The relationship between wire feed speed and voltage for any filler wire type and diameter could be programmed into the power source and a single control could be used to continuously vary mean current. The equipment contained preset resistors to store the fixed parameters but microprocessor-controlled units are now available.

A number of different electronic control systems have been developed for MIG/MAG welding in both the dip and spray transfer modes of operation, which overcome the setting-up difficulties mentioned above.

The setting-up difficulties were particularly acute with the pulsed arc mode of operation, but they have been overcome by the so-called 'synergic control technique' used in conjunction with a transistorised power source. With synergic control, precise independent regulation of the pulse shape, pulse current time, pulse frequency, and background current is obtained. The electronic power source can produce continuously variable 25–250 Hz pulse frequencies. With variable frequency pulsing, the correlation between pulse energy, burn-off rate, and arc characteristics can coincide with all electrode wire feed speeds to provide one metal drop transfer per pulse.[21]

With synergic control, all the pulse parameters are preprogrammed for a wide range of wire feed speeds. During welding the wire feed rate and pulse frequency automatically adjust together to produce one metal droplet transfer at a constant arc length. The welder only needs to adjust one control, i.e. average current.

Modern power supplies have control systems based on microprocessor technology. Memory chips in the control unit

store process data and produce the optimum operating parameters if the user presses the appropriate switches to specify the types of filler wire and shielding gas. In some power sources the user can load his own operating programmes into the control unit. For further information the reader should consult references 22 and 23.

Similarly to manual metal arc welding, there are a number of 'hobby' sets on the market which can be used on the 13 A mains. These have a limited number of current settings for use with 0.6 or 0.8 mm diameter wire and can be used to weld carbon steel, stainless steel and aluminium in thicknesses up to 6 mm. Tuition by video is available.[24]

Applications MIG and MAG welding with the various modes of metal transfer can be used for similar applications to those that are fabricated by manual metal arc (MMA) welding. In addition, they are more suitable for welding some of the non-ferrous alloys such as aluminium and copper alloys and are probably equally suitable for welding stainless steel and nickel alloys. For sheet metal thicknesses which are welded by a single run of weld metal, MIG and MAG welding are generally up to 50% faster than MMA welding. In thicker materials, MIG/MAG welding and MMA welding used with the same duty cycle, i.e. the proportion of arcing time to total time, will have approximately the same overall welding rate, provided that full use is made of positioners to enable a large proportion of the welding to be carried out in the flat or horizontal–vertical positions. Claims made in the literature or in suppliers' brochures about the superiority of one process over the other in respect of productivity and economic advantages should be treated with caution, because they may only be valid for a specific application. One great advantage of MIG/MAG welding over MMA welding is the ease with which the process can be mechanised either by fitting the welding gun to a traverse unit, or by moving the work piece under a stationary gun either by linear motion or rotation.

Robotic and automated MIG/MAG welding has advanced with the developments in microcomputers and electronic power sources, the latter providing very stable arcing conditions in spite of mains voltage fluctuations. Automated MIG/MAG welding utilises seam-tracking devices which are necessary to compensate for inaccuracies of the component parts or distortion during welding.

Seam-tracking devices contain contact type sensors such as probes or guide wheels or the non-contact types such as electromagnetic, ultrasonic or video systems. For further information on robotic or automated welding the reader is referred to references 25–27.

9.2.1.8 *Flux cored arc welding*

Flux cored arc welding is similar in many respects to MIG/MAG welding, except that in one version of the process, no shielding gas is added. In this case the gas shield originates from the decomposition of minerals contained in the tubular cored electrode and this version of the process is sometimes referred to as 'self-shielded welding'.

Cored electrodes in coiled form are manufactured from steel strip which is first bent into a U section as it passes through forming rolls. The U-shaped strip is then filled with a metered quantity of flux and metal powders and the strip is passed through dies to form it into circular cross-section of 0.9–3.2 mm diameter.

Tubular cored electrodes Tubular cored electrodes may be gas shielded with CO_2 or Ar/CO_2 mixtures, or they may be self-shielded. Cored electrodes are classified according to the constituents contained in the core which influence the characteristics of the electrode.

For a full description of the different carbon and carbon–manganese steel types BS 7084: 1989,[28] AWS A5.20-79[29] and AWS A5.29-86[30] should be consulted. Many higher tensile and low alloy steels can be welded with flux-cored wires of matching strengths.

Stainless steel cored wires are available for use either with or without shielding gas, and many different types of cored wires are used for hard-facing applications in which a coating is applied to a steel base to confer resistance to wear, corrosion or heat.

Application Flux cored arc welding is used for similar applications to manual metal arc or MIG/MAG welding and, like MIG/MAG welding, the process can be mechanised.

Tubular cored electrodes are available in a wider range of compositions than are solid wires because of the ease of introducing alloying elements in powder form. Flux cored wires, particularly the gas-shielded types, meet the mechanical-property requirements of a range of applications and some grades give good low temperature impact properties. The mechanical properties attainable with self-shielded cored wires is more limited with maximum weld metal strengths of 700 N mm^{-2}.

Self-shielded wires are particularly useful for site work because no bottles of shielded gases, which are unwieldly, are required. Another advantage on site is that there is no externally added shielding gas which is susceptible to disruption by wind.

Flux cored wires can be used at higher maximum currents than can solid wires, resulting in high deposition rates.

9.2.1.9 *Gas shielding tungsten arc (TIG) welding*

In this process an arc is established between a tungsten electrode and the parent metal forming a weldpool into which filler rod is fed, generally by hand (*Figure 9.22*). Mechanised systems which feed the filler wire are available and movement of the welding head along the joint line can also be mechanised. The tungsten electrode is non-consumable and contamination of the weld pool by air is prevented by an inert shielding gas such as argon, helium, or mixtures of these gases.

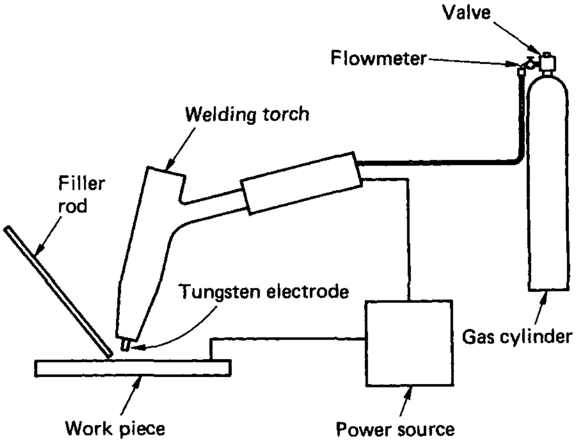


Figure 9.22 Tungsten inert gas (TIG) welding

A high level of skill is required by the welder to control penetration with great precision, which makes the process particularly suitable for the welding of thin sections and for the deposition of root runs in pipe.

Electrodes and filler rods Pure tungsten electrodes can be used, but improved arc initiation and stability are obtained by the use of electrodes containing additions of either thoria (thorium oxide) or zirconia (zirconium oxide). Thoriated electrodes are preferred for d.c. welding and zirconiated electrodes are used for a.c. welding. Electrode diameters vary from 1.2 to 4.8 mm, depending on the welding currents used, which can range from 75 to 450 A for thoriated electrodes and 50 to 200 A for the zirconiated types.

Filler rods which are specified in BS 2901: Parts 1–5: 1992 are of 1.2–5.0 mm diameter and are available in a wide range of compositions suitable for welding carbon and low alloy steels, stainless steels, copper and copper alloys, nickel and nickel alloys, aluminium and aluminium alloys, titanium and zirconium.

Power sources An a.c. or d.c. power source with standard generators, rectifiers or transformers is used. For stable operation the power source must have a 'drooping characteristic' so that when variations occur in voltage or arc length the current remains substantially constant. When changes occur in the arc length in which the welding torch is manually guided along the joint line, the power input remains within ±8% of the preset value.

If the arc is initiated by touching the tungsten electrode onto the parent metal, the electrode becomes contaminated and, to avoid this, a high frequency oscillator is incorporated in the power source. Alternatively, a spark starter using a high voltage coil similar to that in a car ignition circuit can be used. When the gas in the gap between the electrode and the parent plate is ionised by either the high frequency or the spark discharge the full welding current flows. With d.c., the high frequency is normally turned off automatically after arc initiation, but with a.c. it is operated continuously to maintain ionisation of the arc path when the arc voltage passes through zero.

Power sources are available for pulsed arc welding which enables a stable arc to be maintained at low currents down to 10 A. In pulsed TIG welding, the pulse frequency varies from 10 per second to 1 per second and each pulse forms a molten pool which solidifies before the next pulse. Pulsed TIG welding can be used to control penetration in thin sheet and in the root runs of pipes and positional welds in plate.

Applications TIG welding is particularly suited to welding light gauge carbon, alloy and stainless steels and all non-ferrous metals and alloys. A clear clean weld pool is formed with precise control of heat input and the ability to weld with or without filler metal in all positions makes the process attractive for critical applications where exceptionally high quality is essential. Examples are stainless steel piping for nuclear applications and the wide range of piping compositions used in chemical plant.

For such critical applications, fully mechanised orbital welding equipment has been developed in which the welding torch and wire feeding mechanism rotates round the pipe joint. Thin and thick section pipes can be welded with a narrow gap joint preparation and *in situ* fabrication of nuclear and chemical plant is now possible.

Other specialised TIG welding equipment is used for the mechanised welding of tubes to tube plates.

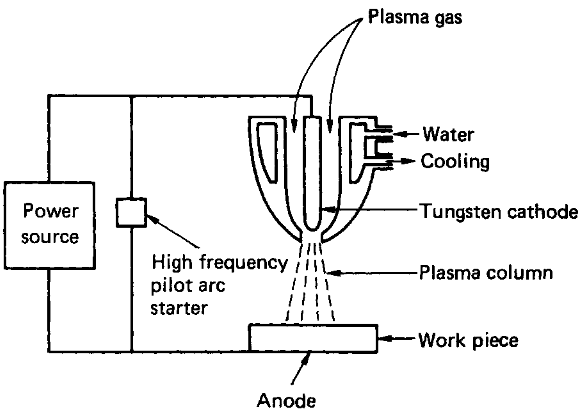


Figure 9.23 Plasma arc welding

9.2.1.10 Plasma arc welding and cutting

Plasma arc welding was developed from TIG welding by placing a narrow orifice round the arc and supplying a small flow of argon through the orifice (*Figure 9.23*). The constricted arc dissociates the argon gas into positive and negatively charged electrons to form a plasma. When the plasma gas flows away from the arc column, it forms neutral atoms again and gives up its energy in the form of heat.

A low current pilot arc is initiated between the tungsten electrode and the water cooled copper orifice. The argon gas flowing through the orifice is ionised and initiates the primary arc between the tungsten electrode and the parent metal when the current is increased.

The arc and the weld zone are shielded by a gas flowing through an outer nozzle. The shielding gas consists of argon, helium or gas mixtures of argon with either hydrogen or helium.

A normal tungsten arc has a temperature of approximately 11 000°C, but the constricted arc of a plasma torch can reach 20 000°C. The high temperature ionised gas jet gives up its energy when it contacts the parent metal and thus increases the energy of the tungsten arc.

This produces a deep penetration weld with a high depth-to-width ratio with minimum distortion of the parent metal. The term 'keyhole' is used to describe the shape of the hole formed in the parent metal when a close square edge butt joint is welded. As the torch is moved along the joint molten metal flows round the edges of the hole and solidifies at the rear of the hole. The molten metal at the sides of the hole is held in place by surface tension and the pressure of metal vapour in the hole.

The keyholing welding technique can be used on carbon, low alloy steels and stainless steels in thicknesses of 2.5–10 mm and in aluminium alloys up to 20 mm. Welding speeds are generally 50–150% higher than those possible with TIG welding.

A low current version of the process is microplasma arc welding which is used for precision welding of thin sheet (0.025–1.5 mm thick) at currents of 0.1–10 A. The plasma arc is much more stable than a TIG arc, which tends to wander from the joint line at low currents.

Plasma cutting If the current and gas flows are increased sufficiently the molten metal formed round the keyhole is

ejected at the bottom of the hole and, as the plasma torch is traversed along the work piece, a cut is formed. Plasma cutting is especially suitable for cutting non-ferrous metals, such as aluminium, copper, and nickel and their alloys, which are not easily cut by oxy-fuel gas flames. Most non-ferrous metals are cut using nitrogen, nitrogen–hydrogen mixtures or argon–hydrogen mixtures as the plasma gas. A secondary shielding gas delivered through a nozzle that encircles the plasma gas nozzle is selected according to the material being cut. For mild steel and stainless steel it can be CO_2, and for aluminium it is an argon–hydrogen mixture. Sometimes water is used in place of the ancillary shielding gas and in another variety of the process water is injected round the end of the plasma gas nozzle which has the effect of concentrating the plasma flame and allowing higher cutting speeds.

Plasma cutting can be used for plate-edge preparation, i.e. bevelling and for shape cutting. The process can be used manually or the torch can be mounted on mechanised cutting equipment identical to that used for oxy-fuel gas cutting. For metal thicknesses up to 75 mm, carbon steels can be cut faster by plasma cutting than by oxy-fuel gas, and up to 25 mm thick the cutting speeds can be five times as fast.

An important variation of the process involves the use of compressed air for the plasma gas without the provision of any additional shielding gas. The use of compressed air in place of water for cooling enables the torch to be of simplified construction.

Small manual air plasma torches are available which are finding increasing application in sheet-metal cutting, e.g. motor repair shops.

For further information the reader is referred to reference 31.

9.2.1.11 *Gas welding and cutting*

Gas welding is carried out using a flame produced by burning approximately equal volumes of oxygen and acetylene which are delivered at equal pressures from gas bottles to a welding torch. The flame temperature is approximately 3100°C, which is high enough to melt steel and other metals. Filler metal, if required, is added by manually feeding a rod into the front edge of the weld pool while the torch is moved along the joint. The products of combustion provide sufficient protection from the atmosphere when welding steel. When welding other metals such as cast iron, stainless steel, aluminium alloys and copper alloys, fluxes are used to clean and protect the metal from oxidation.

Equipment The welding torch has two knurled control knobs which regulate the flow rates of oxygen and acetylene so that a neutral or slightly oxidising or reducing flame is obtained depending on the application. The torch has a screw-in nozzle from a set of nozzles having different diameter holes which produce the appropriate size of flame and, therefore, the required heat input for the particular metal and thickness to be welded.

The oxygen and fuel gas hoses are connected between the welding torch and the gas bottles, the gases passing through flash-back arresters and pressure regulators. Flash-back arresters are safety devices that prevent a flame travelling back into the cylinders in case a backfire occurs.

For workshop use, the gas bottles are generally mounted in pairs on a trolley which can be moved to the place where it is required.

Filler metal and fluxes Chemical compositions of filler metals are specified in BS 1453: 1972,[32] which includes ferritic steels, cast iron, austenitic stainless steels, copper and copper alloys and aluminium alloys.

Ferritic steels do not require the use of a flux; proprietary fluxes are available for other materials.

Applications Gas welding is used mainly for repair and maintenance work, particularly in the repair of car bodies and agricultural implements, although it is slowly being replaced by small TIG and MIG welding equipment.

Gas welding is used to a certain extent for sheet-metal work, i.e. heating and ventilating ducting, and is still used for making the root runs in pipes where it is particularly useful for bridging gaps.

Two applications where gas welding has distinct advantages over other processes are in the welding and repair of grey iron castings and in hardfacing with expensive alloys.

Grey iron castings can be successfully welded by the use of high preheating temperatures of up to 600°C and gas welding with cast iron filler rods.

The deposition of high cost wear-resistant alloys such as the cobalt–chromium tungsten types or those based on chromium or tungsten carbides, can be carried out with minimum melting of the parent metal so that dilution of the deposited alloy and the consequent decrease in wear resistance is avoided.

Gas welding is also successfully applied in jewellery manufacturing with miniature torches and small gas bottles.

Acetylene is the only fuel gas suitable for gas welding because of its favourable flame characteristics of both high temperature and high propagation rates. Other fuel gases, such as propane, propylene, or natural gas produce insufficient heat input for welding but are used for cutting, torch brazing and soldering. They are also used for flame straightening of distorted components and for preheating before welding and post-heating after welding.

Gas cutting Gas cutting, sometimes referred to as 'flame cutting' or 'oxygen cutting', involves an active exothermic oxidation of the steel being cut when the material has been preheated by an oxy-fuel gas flame to the ignition temperature of around 900°C.

The equipment used for gas cutting is the same as that used for welding except that a special cutting nozzle is required. The nozzle has an outer ring of holes through which the preheating gas mixture is delivered and a central hole through which the oxygen jet flows. The exothermic reaction of oxidation of steel forms a fluid slag of iron oxide and, after a few seconds (depending on the metal thickness), the section is pierced. Iron oxide and molten metal are expelled from the cut by the oxygen stream. Movement of the cutting torch across the work piece produces a continuous cutting action and the torch can be operated manually or by a motorised carriage. Steel up to 300 mm thick can be cut by this process.

Oxidation resistant steels such as stainless steels may be cut by specialised methods including the introduction of iron powder or other proprietary powders into the oxygen stream. These powders react with the refractory chromium oxides and reduce their melting points and increase their fluidity, enabling cutting to take place. For further information the reader should consult reference 31.

Stainless steels and non-ferrous metals and alloys are usually cut by using the plasma cutting process which does not rely on an exothermic reaction.

Manual gas cutting is possible and the accuracy of cutting can be improved by the use of a small wheel mounted on the cutting torch. The wheel may be free running or motorised. For general cutting and profiling, including cutting bevels on plate edges, mechanised cutting is normally used. For mechanised cutting, electronic tracing devices are commonly used.

These consist of a photoelectric cell which follows the outline of a drawing and guides the cutting nozzle by means of driving motors which regulate the movement of a carriage and cross arm to which the torch is attached.

Numerically controlled cutting machines are available which use programs stored or punched on magnetic tape which send appropriate signals to the drive motors.

9.2.1.12 *Welding and cutting with power beams*

Electron beam welding and laser welding utilise high-energy beams which are focused on a spot on the work-piece surface about 0.2 mm in diameter. This intense heat source, which releases its kinetic energy when the beam hits the surface, is radically different from arc welding in which the arc melts an area of about 5–20 mm diameter, depending on the welding conditions.

When the power density of an electron or a laser beam at the focused spot is 10 kW mm^{-2} or greater, energy is delivered at a faster rate than can be conducted away in the form of heat in the work piece and the progressive vaporisation of metal through the section thickness forms a hole. If the beam is then traversed along the work piece, molten metal flows around the sides of the hole and solidifies at the rear of the hole. Molten metal at the sides of the hole is held in place by surface tension and the presence of metal vapour in the hole in the same manner as that described for the keyholing technique in Section 9.2.1.10 for making a square edge butt joint.

As with plasma welding, a deep penetration weld with a high depth-to-width ratio is formed with minimum distortion of the parent metal. Maximum penetration depths in steel are approximately 280 mm for electron-beam welding and 12 mm for laser welding, although developments in the latter process are likely to increase this to 25 mm or more.

Electron-beam welding In this process, a finely focused beam of electrons passes from a cathode and travels through a hole in an anode and is focused on a spot on the work piece 0.2–1 mm in diameter by means of a magnetic lens. Deflection coils are used to cause the beam to move in a circular pattern to increase the width of the weld, so that fusing two mating surfaces together in a close square butt joint is possible. The cathode in the electron gun is maintained at a negative potential of 60–150 kV and the gun is contained in a vacuum of 5×10^{-5} T. In the work chamber, a pressure of 5×18^{-3} T is suitable for welding most metals and for some applications a pressure of 10^{-1} T or less is used with the advantage of much shorter chamber excavation times resulting in increased production rates.

High vacuum welding High vacuum welding, in which the work chamber is maintained at a pressure of 10^{-3} to 10^{-5} T, depending on the application, has the following features.

(1) Maximum weld penetration and minimum weld width and shrinkage, enabling all thicknesses to be welded in a single pass.
(2) The highest purity weld metal is produced because of the absence of any contaminating gases such as oxygen or nitrogen.
(3) A high vacuum allows a long distance to be maintained between the gun and the work piece, which facilitates observation of the welding process.
(4) The pump-down time is lengthy (up to 1 h or more, depending on the size of the chamber) which lowers the production rate of small jobs with shallow welds but is insignificant when welds in plate of 50 mm or more thickness are made with a single pass.

Welding speeds for steel are shown in *Table 9.1*.

Medium vacuum welding In medium vacuum welding the working chamber is maintained at a pressure within the range 10^{-3} to 10^{-1} T, although pressures of up to 25 T are reported in the literature[33] together with correspondingly short pump-down times of a few seconds.

Medium vacuum welding with small working chambers is used extensively for small repetition work such as welding of finish machined gear trains and similar high-volume mass-production applications for the motor industry. Many high precision semi-finished or fully machined components for aircraft engines are also welded by using the electron beam process.

Medium vacuum electron-beam welding is not suitable for welding reactive metals and alloys such as titanium and zirconium which require the high vacuum process to obtain sound welds.

Out of vacuum electron-beam welding Provided that the gun-to-work distance is less than about 35 mm, so that the greater dispersion of the electron beam compared with working in a vacuum is allowed for, it is possible to weld many materials out of vacuum. With 60 kW non-vacuum equipment, single-pass welds can be made in metal thicknesses up to 25 mm. Metals welded out of vacuum include carbon and low alloy steels, and copper and aluminium alloys. Because of the presence of air, which can cause contamination of the weld metal, it is usually necessary to provide an inert shielding gas to cover the weld zone.

Laser welding and cutting Two types of laser are used for welding and cutting: the solid state yttrium–aluminium–garnet (YAG) laser, and the carbon dioxide (CO_2) laser.

The term 'laser' is an acronym for 'light amplification by stimulated emission of radiation'. A laser is a device for producing monochromatic (single wavelength) light which is coherent, i.e. all the waves are in the same phase. A laser beam can be transmitted over a distance of many metres and can be focused to produce the high-energy density required for welding or cutting.

The solid state YAG laser The solid state laser is stimulated to emit coherent radiation by means of the light from one or more powerful flash tubes. The output is in the infra-red region (around 1.06 μm). Both input and output are generally pulsed, and the maximum power output is 500 W.

Pulsed laser welding is used extensively in the electronics industry where miniaturisation requires very precise position-

Table 9.1 Penetration depth and welding rates

Welding power (kW)	*Plate thickness* (mm)	*Welding speed* ($mm\ min^{-1}$)
0.25	1	1500
2.0	10	750
4.5	10	750
15	50	150
75	280	50

ing of small welds. Typical examples are encapsulation of microelectronic packages, and the joining of fine wires by butt, lapped, or cross-wire joints. Spot welds can be made between overlapping sheets and seam welds can be formed by a series of overlapping spot welds. Most metals can be welded, including steel, copper, nickel, aluminium, titanium, niobium, tantalum, and their alloys.

Solid state lasers are also used for precision drilling of holes having very small diameters.

The carbon dioxide laser In the carbon dioxide laser the lasing medium is a gaseous mixture of carbon dioxide, nitrogen and helium at a reduced pressure of 2–50 T. The output of the carbon dioxide laser is in the mid-infra-red region (10.6 μm) and CO_2 lasers are available having powers up to 20 kW. The high power density produced (10^4 W mm^{-2} or more) forms a cavity or keyhole in the work piece which enables a deep penetration weld to be produced that is similar in appearance to an electron-beam weld.

Penetration is less than for an electron beam operating at the same power because of the plasma formed in the keyhole. The plasma gas escapes from the keyhole and interacts with the laser beam, restricting penetration to approximately 12 mm in mild steel. Methods have been developed for overcoming this problem,[33] either by the use of a pulsed output from the laser with pulses of shorter time than that required to generate the plasma, or by the use of a high velocity jet of helium to disrupt the plasma above the weld. Such developments are likely to increase the thicknesses that can be laser welded to 25 mm or more.

Carbon dioxide lasers are used in production in some automotive and aerospace applications, particularly for titanium and nickel alloys. The advantage of laser welding over electron-beam welding is that a vacuum is not required, which simplifies the welding operation. Other materials that can be welded using the carbon dioxide laser are steels, copper alloys, zirconium and refractory metals, but aluminium alloys are not readily weldable with this laser because of their reflectivity of the laser beam.

Laser cutting Cutting with the carbon dioxide laser is carried out by a combination of melting and vaporisation, with an auxiliary jet of gas to blow the molten metal from the cut. Various gases are used for this purpose, including oxygen, compressed air, inert gases, and carbon dioxide.

Oxygen produces an exothermic reaction with ferrous metals which increases the efficiency of cutting. The inert gases produce clean unoxidised surfaces, which are important when cutting readily oxidised metals such as aluminium and titanium.

9.2.1.13 Resistance welding

Spot, seam and projection welding are carried out by electric resistance heating of two overlapping metal parts which are pressed together by copper or copper alloy electrodes. Local melting occurs at the faying surfaces and an internal weld nugget is formed.

The welding cycle comprising current, pressure and time is readily controlled automatically, giving the following advantages:

(1) little skill is required for the operation;
(2) welding can be readily built into production lines;
(3) welding can be associated with automatic loading, unloading and transferring of components, so that resistance welding is the simplest process for automation or robotic welding;
(4) welding times are short so that output is high;
(5) no filler wires or fluxes are used; and
(6) distortion is a minimum, as heating is confined to a small area.

The high currents required for resistance welding are generally obtained from a single-phase a.c. transformer having a primary winding of several hundred turns. The secondary winding consists of one or two turns of thick copper which may be water cooled. The voltage is stepped down to a value between 4 and 20 V.

Three-phase welding machines are available; these are more expensive but have a better power factor. A three-phase transformer is sometimes used with a d.c. rectifier to carry out resistance welding with d.c.

For further information on power sources the reader should consult reference 33.

Spot welding Spot welding (*Figure 9.24*) is used for the fabrication of sheet metal of 0.6–3 mm thickness to produce lap joints that are intermittently welded and, therefore, not pressure tight. Typical applications are low carbon steel components for car bodies, cabinets, and general sheet-metal work. Spot welding is also applicable to stainless steel, aluminium and aluminium alloys, and copper alloys.

Because of its high electrical conductivity, pure aluminium is difficult to spot weld, and its softness results in heavy indentation by the electrodes. Similarly, the high electrical conductivity of copper makes it unsuitable for spot welding. Stainless and heat-resisting steels are spot welded for aircraft and gas turbine engines.

Seam welding Seam welding (*Figure 9.25*) is similar in principle to spot welding and uses rotating wheel electrodes which roll the overlapping components between them. The welding current passes intermittently through the electrodes forming a series of welds that overlap one another. The electrode wheels rotate continuously, and at least one is power driven to move the component along, in addition to carrying the welding current.

Pressure-tight seams can readily be made and the process is faster than spot welding. However, machines used for seam welding are heavier and more costly than spot welding machines. Seam welding is primarily suited to making long straight welds, although curved welds can be made; for example, welds which may occur at the corners when joining two half pressings together to form a fuel tank.

Typical applications are pressure-tight seams for oil drums and refrigerator parts.

Projection welding In this process the current is concentrated by the shape of the components themselves, small dimples

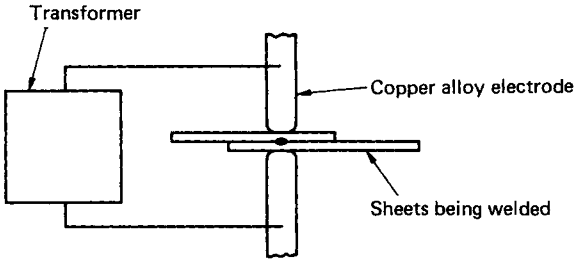


Figure 9.24 Resistance spot welding

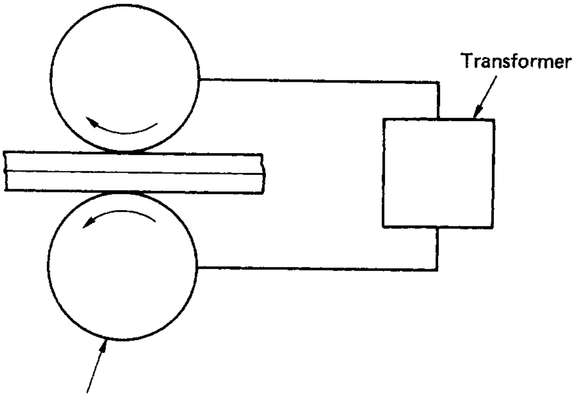


Figure 9.25 Resistance seam welding

being formed on one of the sheets to be joined, as shown diagrammatically in *Figure 9.26*.

The electrodes are of relatively large area compared with those used in spot welding and can, therefore, be made of a hard material having a comparatively high electrical resistance, such as a copper–tungsten alloy. As several welds can be made at the same time, the welding machine must be of high electrical capacity and must also be capable of applying the high total mechanical load to the components during welding. Three projections welded simultaneously give the best results, and up to five projections are common. Applications involving large numbers of projections are known, but as the number increases it becomes more and more difficult to ensure uniform and adequate welding of all the projections. The process can also be used for attaching studs, nuts and discs to flat plates.

The main features of the process are listed below.

(1) Several welds can be made simultaneously in the time it would take to make a single spot weld.
(2) Electrodes have longer life between dressings compared with spot welding electrodes, because of the relatively low current densities used.
(3) Welds are generally of better external appearance than spot welds due to the absence of indentations.
(4) T joints may be made as well as lap joints.
(5) Machines are generally more costly than spot welding machines.
(6) The cost of preparing components is greater because of the necessity of forming projections.

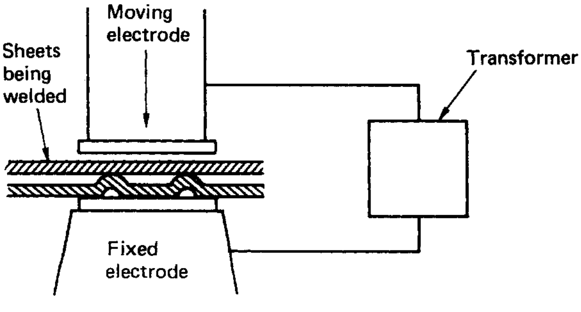


Figure 9.26 Projection welding

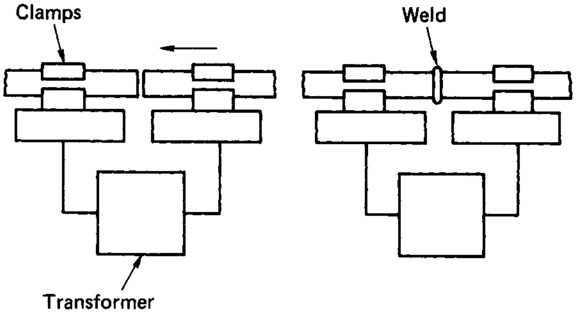


Figure 9.27 Resistance butt welding

9.2.1.14 Resistance butt and flash welding

Resistance butt welding The two ends of the parts to be joined are brought into contact and current is passed across the joint while a moderate mechanical force is applied to the components (*Figure 9.27*). As the joint is heated up by electrical resistance, the material softens and welding takes place. The finished joint shows a thickening or upsetting of the components in the joint area, and for some distance away from the joint. Resistance butt welding is essentially a solid state welding process in which no melting occurs.

Recrystallisation of the metal takes place across the faying surfaces, and the upsetting action has the effect of removing oxides from the joint.

The process has two main applications:

(1) joining two components of the same cross-section end to end, e.g. wire, rod, bar or tubing, as used in rod and wire mills for joining the ends of coils for continuous processing; and
(2) continuous welding of longitudinal seams in pipe or tubing formed from flat plate.

The process is applicable to the welding of carbon, alloy and stainless steel, aluminium alloys, copper alloys, and nickel alloys, and electrical resistance alloys.

Resistance butt welding machines are generally designed to weld a particular family of alloys because the current densities and pressures as well as the rate of application of the pressure differ widely between, for example, steels and aluminium alloys. The cross-sectional areas to be joined also determine the current capacity of the welding transformer.

The welding machine has two platens on which are mounted clamping dies which grip the components. One platen is stationary and the other moves to produce the pressure required.

Flash welding Flash welding was developed from resistance butt welding; the principle is shown in *Figure 9.28*. The current is switched on before the ends of the components are in contact, with the result that small volumes of metal melt explosively at the points where contact is first made. This process continues as the two components are moved towards one another, causing a large amount of flashing and removal of metal, while at the same time heating the areas in contact and the material immediately behind. Once suitable temperature conditions have been established, the two ends of the components are forced together by a sudden increase in the

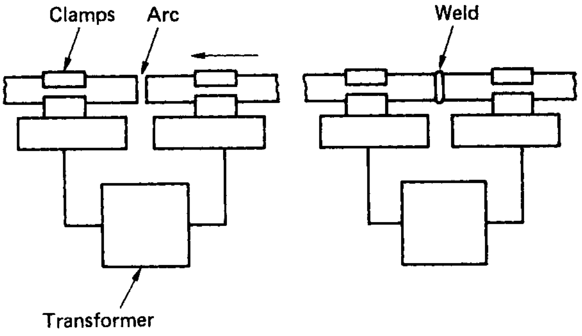


Figure 9.28 Flash welding

load applied and the current is switched off. The effect of this sudden load is to squeeze out from the joint all the overheated and oxidised metal and to form a high quality pressure weld. The fin or flash of upset metal surrounding the joint is normally removed before the component is put into service.

Typical applications of flash welding are the production of continuous welded railway track, motor car wheel rims formed from flat steel stock and mitre joints in door and window frames. The materials that can be welded using this process are similar to those that can be resistance butt welded.

9.2.1.15 Friction welding

Friction welding is a solid state joining process, i.e. there is no fused metal involved. The heat required to make a forge weld is produced by moving one component relative to a mating component under pressure. The motion is usually rotational, although linear relative motion can be used and is under development.

A friction weld is similar in appearance to a resistance butt or a flash weld, with upset metal or flash which is generally removed after welding by a machining operation.

Butt welds can be made in rods or tubes, and rods or tubes can be welded to plates. One of the components is generally circular in cross-section, although square rods can be joined by friction welding. A wide range of sizes can be friction welded from small-diameter wires used in the electronics industry up to 150 mm diameter aluminium bus-bars.

Typical applications are found where high production rates are required, e.g. in the motor industry. Examples are axles, drive shafts, steering shafts and valves where a high alloy heat-resisting head is welded to a cheaper carbon steel shank. Bar stock can be welded to plates to produce parts that would normally be forged.

Most metals and alloys, with the notable exception of cast iron, can be friction welded. Dissimilar metals can also be welded without the formation of a brittle zone, which often occurs with arc welding, e.g. aluminium can be welded to steel.

Friction surfacing is an important development in which a corrosion or wear resistant alloy rod is rotated rapidly under pressure against a surface to be clad with the alloy. Under carefully controlled conditions a friction weld can be deposited over the surface and the weld metal is unchanged in composition because there is no dilution of the weld caused by melting of the parent metal as occurs in surfacing by arc welding. For further information on friction welding the reader is referred to reference 33.

9.2.2 Soldering and brazing

Brazing and soldering are carried out at temperatures below the melting points of the metals being joined. Brazing filler metals have melting points above 450°C and solders melt below this temperature. In both processes the molten filler metal flows by capillary action between closely fitted surfaces of the parts to be joined.

Unlike welding, soldering and brazing do not generally produce joints having mechanical properties matching those of the parent metal.

Typical applications, such as car radiators and electrical connections, require leak-tightness and good electrical conductivity, respectively. If high strength or ductility is required, correct alloy selection and joint design are essential. For further guidance on this subject references 34 and 35 should be consulted.

The low temperatures involved in soldering and brazing have two beneficial effects compared with welding: (1) parent metal properties are less affected, especially by soldering; and (2) residual stresses and distortion are lower.

For successful joining, the following procedures are required.

(1) For wetting of the metal by the solder or brazing alloy, the surfaces must be chemically clean. Degreasing and mechanical cleaning or pickling may be necessary.
(2) During the joining operation a flux is generally used to remove surface oxides and to prevent them from reforming. In some brazing processes a reducing or neutral atmosphere or a vacuum is used.
(3) Heating of the parent metal must be carefully controlled so that both mating surfaces reach the melting temperature of the filler metal before it is applied to the joint. When the filler metal is preplaced in the joint the same requirement applies because the molten filler metal will not wet the surface unless the surface is hot enough.
(4) All traces of flux residues which might lead to corrosion must be removed after soldering or brazing.

9.2.2.1 Soldering

Solders The normal range of soft solders is listed in BS 219: 1977,[36] which also indicates typical uses. The commonest solders are the tin–lead alloys, some of which have antimony added. These solders are used in various applications, depending on the composition, including manufacture of food-handling equipment, tin-plated cans, domestic utensils, electronic assemblies, heat exchangers, and general engineering work. Tin–antimony solder is used for higher service temperatures, e.g. above 100°C, and in applications for joining stainless steels where lead contamination must be avoided. The alloy is also used for plumbing and refrigeration applications.

Tin–silver solder is used for fine instrument work and food applications while tin–lead–silver solder is used where service temperatures will be above 100°C or below −60°C.

Soldering methods Many sources of heat can be used for soldering, including heated iron, flame or torch, hot-plate or oven, high-frequency induction, electric resistance, hot gas, as well as dip, wave or flow methods involving molten solder baths.

For further information on both manual and mechanised soldering techniques the reader should consult references 35, 37 and 38.

Manual soldering methods involve the use of a soldering iron or a torch, are fairly slow and require a reasonable degree of skill, although an experienced operator can achieve high production rates on repetitive work. Dip soldering is faster

and the electronics industry uses specially designed equipment for wave soldering of printed circuit boards.

9.2.2.2 Brazing

Brazing alloys are specified in BS 1845: 1984[39] divided into eight groups based on aluminium, silver, copper–phosphorus, copper, copper–zinc, nickel, palladium and gold. The choice of brazing alloy for a particular parent metal can be fairly wide and a decision may be based on metallurgical considerations including corrosion resistance, the maximum temperature to which the parent metal can be heated to avoid deterioration in properties, or melting of previously applied brazing alloys. For further information the reader should consult references 34, 40 and 41.

A brief summary of the uses of the eight main groups of brazing alloys is given in *Table 9.2*.

Brazing alloys are available in the form of wires, rods, powders and inserts of various shapes for preplacing in the joint.

Heating methods A full description of the various heating methods and their advantages and disadvantages are described in BS 1723: 1986: Part 2.[40] The methods comprise hand torch brazing, mechanised flame brazing, induction brazing, furnace brazing including protective atmosphere, vacuum and open furnace brazing, and immersion brazing, including flux bath, dip bath and salt bath brazing. Special processes referred to are infra-red brazing and laser brazing. Reference 40 also contains details of design and location methods for brazing.

Applications Brazing is applicable to cast irons, steels, galvanised steel, aluminium, copper, magnesium, nickel and their alloys, stainless and heat-resisting steels, titanium, zirconium, ceramics and refractory metals and to dissimilar metal joints.

Brazing is selected in preference to other joining processes in the following cases:

(1) where heating must be restricted to avoid melting or distortion of one of the parts to be joined;
(2) where strength or dimensional accuracy of the assembly would be impaired by heating to high temperature;
(3) where a soft soldered joint would not be strong enough;
(4) where parts cannot be welded because of the properties of the parent materials involved, e.g. ceramics and refractory metals;
(5) where a number of joints must be made in a small complicated assembly; and
(6) where joints would be inaccessible to the welding processes.

9.2.3 Productivity and welding economics

Productivity, in the broadest sense, is what determines profit. Unfortunately, the cost of welding is one of the least documented topics in the whole field of welded fabrication. When a company decides to review its fabrication methods there is often a lack of understanding of the principles involved and only limited knowledge of the alternatives available. Reliance may be placed on equipment salesmen who may be fair and have an excellent product to sell but, quite naturally, may be biased in their approach.

There is a lack of information on welding compared with that readily available on machine tools, e.g. cutting speeds and depth of cut, which can enable a machine to be integrated into a manufacturing system to give a high utilisation or duty cycle and, therefore, maximise productivity. The same objective should be applied to welding as part of the manufacturing system.

The first task in reviewing fabrication methods is a simple one and concerns design. In the case of arc welding, fabrications should be designed to have the minimum amount of weld metal that will meet the service requirements of the components in terms of mechanical proeprties.[42] The total quantity of weld metal in a fabrication is, of course, only a very small proportion of the total weight (perhaps 2% or less), but nevertheless it is sensible to use the minimum amount necessary. This requires close attention to detail in design, shop floor supervision and inspection to ensure that welded joints are not overdesigned and weld metal is not being wasted.

Sometimes partial penetration butt joints will be adequate and intermittent fillet welds can replace continuous ones. Fillet weld profiles should not be excessively convex because

Table 9.2 Main uses of brazing alloys

Brazing alloy	*Materials suitable for brazing*
Aluminium	Pure aluminium and some aluminium alloys
Silver	Most ferrous and non-ferrous metals and alloys, except aluminium, magnesium and refractory metals and their alloys
Copper–phosphorus	Copper and copper alloys
Copper	Ferrous materials in a protective atmosphere or in a vacuum
Copper–zinc	Stainless steel and other heat- and corrosion-resistant alloys in a protective atmosphere or in a vacuum
Nickel	Stainless steel and nickel–chromium heat resisting alloys for high-temperature service
Palladium and gold	Metallised ceramics, copper, nickel and ferrous alloys in a protective atmosphere or in a vacuum

10

Electrical and Electronics Principles

C J Fraser

Contents

10.1 Introduction 10/3

10.2 Basic electrical technology 10/3
- 10.2.1 Flux and potential difference 10/3
- 10.2.2 Simple resistive circuits 10/3
- 10.2.3 Electromotive force and potential difference 10/4
- 10.2.4 Power and energy 10/4
- 10.2.5 Network theorems 10/4
- 10.2.6 Double-subscript notation 10/4
- 10.2.7 Electrostatic systems 10/4
- 10.2.8 Simple capacitive circuits 10/5
- 10.2.9 Charging a capacitor 10/6
- 10.2.10 Types of capacitor 10/6
- 10.2.11 Dielectric strength 10/6
- 10.2.12 Electromagnetic systems 10/6
- 10.2.13 Magnetic field of a toroid 10/7
- 10.2.14 Permeability 10/7
- 10.2.15 Faraday's law 10/7
- 10.2.16 Self-induced e.m.f. 10/8
- 10.2.17 Energy stored in an inductor 10/9
- 10.2.18 Mutual inductance 10/9
- 10.2.19 Hysteresis in magnetic circuits 10/9
- 10.2.20 Eddy current loss 10/10
- 10.2.21 Kirchhoff's laws and the magnetic circuit 10/10
- 10.2.22 Alternating quantities 10/11
- 10.2.23 Relationships between voltage and current in resistive, inductive and capacitive elements 10/11
- 10.2.24 Resistive–inductive and resistive–capacitance circuits under transient switching conditions 10/11
- 10.2.25 Steady-state a.c. 10/13
- 10.2.26 Phasor diagrams 10/14
- 10.2.27 Complex notation 10/14
- 10.2.28 The parallel resistive–inductive–capacitance circuit 10/14
- 10.2.29 Power and power factor in a.c. circuits 10/15
- 10.2.30 Frequency response of circuits 10/15
- 10.2.31 Semiconductors 10/16
- 10.2.32 Doped semiconductors 10/16
- 10.2.33 p–n Junction diode 10/17
- 10.2.34 A.c. rectification 10/17
- 10.2.35 The Zener diode 10/18

10.3 Analogue and digital electronics theory 10/18
- 10.3.1 The bipolar, or junction, transistor 10/18
- 10.3.2 Common base characteristics 10/19
- 10.3.3 Common emitter characteristics 10/19
- 10.3.4 The transistor in a circuit 10/21
- 10.3.5 The field effect transistor 10/21
- 10.3.6 Integrated circuits 10/22
- 10.3.7 The thyristor 10/22
- 10.3.8 The triac 10/23
- 10.3.9 Amplifiers 10/23
- 10.3.10 Effect of feedback on amplifiers 10/24
- 10.3.11 Noise and distortion in amplifiers 10/24
- 10.3.12 Amplifier frequency response 10/24
- 10.3.13 Positive feedback and stability 10/24
- 10.3.14 The operational amplifier 10/25
- 10.3.15 The differential amplifier 10/27
- 10.3.16 Instrumentation amplifier 10/27
- 10.3.17 Power supplies 10/27
- 10.3.18 Analogue and digital systems 10/29
- 10.3.19 Boolean algebra 10/29
- 10.3.20 Digital electronic gates 10/30
- 10.3.21 Gate symbols 10/31
- 10.3.22 Logic systems using simple gates 10/31
- 10.3.23 Logic systems using NAND and NOR gates only 10/32
- 10.3.24 Unused inputs 10/35
- 10.3.25 Latches 10/35
- 10.3.26 The Karnaugh map 10/36
- 10.3.27 Minimisation of Boolean expressions 10/37
- 10.3.28 Positive and negative logic 10/38
- 10.3.29 Tri-state logic 10/38
- 10.3.30 Sequential logic circuits 10/39
- 10.3.31 Registers and counters 10/40
- 10.3.32 Timers and pulse circuits 10/42
- 10.3.33 Digital computers and microprocessors 10/42
- 10.3.34 Application specific integrated circuits 10/43

10.4 Electrical machines 10/43
- 10.4.1 The d.c. generator 10/43
- 10.4.2 Methods of connection 10/45
- 10.4.3 The separately excited generator 10/45
- 10.4.4 The shunt-wound generator 10/47
- 10.4.5 The series-wound generator 10/47
- 10.4.6 The compound-wound generator 10/47
- 10.4.7 The d.c. motor 10/48
- 10.4.8 The shunt-wound motor 10/48
- 10.4.9 The series-wound motor 10/49
- 10.4.10 The compound-wound motor 10/49
- 10.4.11 Starting d.c. motors 10/49

10.4.12 Speed control of d.c. motors 10/50
10.4.13 Efficiency of d.c. machines 10/51
10.4.14 Three-phase circuits 10/51
10.4.15 Generation of three-phase e.m.f.s 10/51
10.4.16 Star and delta connections 10/51
10.4.17 Three-phase voltage and current relationships 10/52
10.4.18 Power in three-phase circuits 10/52
10.4.19 Three-phase alternators 10/53
10.4.20 Synchronous motors 10/53
10.4.21 Induction motors 10/54
10.4.22 Starting induction motors 10/54
10.4.23 Braking induction motors 10/55
10.4.24 Speed control of induction motors 10/55
10.4.25 Single-phase induction motors 10/56
10.4.26 The d.c. permanent magnet motor 10/57
10.4.27 The stepper motor 10/57
10.4.28 Brushless d.c. motors 10/59
10.4.29 Transformers 10/59
10.4.30 Basic transformer action 10/59
10.4.31 Transformer voltage equation 10/60
10.4.32 Transformer losses 10/60
10.4.33 Determination of transformer losses 10/60
10.4.34 Referred values 10/60
10.4.35 Transformer efficiency 10/60
10.4.36 Voltage regulation 10/61
10.4.37 Three-phase transformers 10/61
10.4.38 Autotransformers 10/61

10.5 Electrical safety 10/62
10.5.1 Electric shock 10/62
10.5.2 Electric burn 10/62
10.5.3 Rescue 10/62
10.5.4 Protection 10/62
10.5.5 Earthing 10/63
10.5.6 Double insulation 10/63
10.5.7 Low-voltage supplies 10/63

10.1 Introduction

Within the confines of this chapter, it is impossible to include elaborate detail on the principles of electrical and electronics technology. A thorough depth of understanding can only come through numerous applications of the basic principles to the solution of practical problems. Since this is beyond the scope of this chapter, the content has been restrictively selected as appropriate to the practical requirements of manufacturing engineers. All the relevant basic principles are stated as matters of fact. For a more penetrating discourse readers must refer to the many standard texts on electrical and electronics technology which are readily available. An adequate number of references are cited for this purpose.

The graphical symbols used in the text are in accordance with BS 3939: 1986: Part 1.[1] Other symbols and abbreviations are consistent with the SI system of units as defined in BS PD 5686: 1972.[2]

10.2 Basic electrical technology

10.2.1 Flux and potential difference

The concept of flux and potential difference enables a unified approach to be adopted for virtually all 'field' type problems. In general, the flowing quantity is termed the 'flux' and the quantity which drives the flow is called the 'potential difference'. This consistency of method is equally applicable to problems in fluid flow, heat transfer, electrical conduction, electrostatics and electromagnetism, to name but a few. It is not particularly surprising, therefore, that thermodynamicists use an electrical analogy to describe many heat conduction phenomena, while electrical engineers often use heat and fluid flow analogies to enhance an understanding of electrical and electronics principles.

In general terms, the flux may be written as:

$$\text{Flux} = \frac{\text{Field characteristic} \times \text{Cross-sectional area} \times \text{Potential difference}}{\text{Length}} \quad (10.1)$$

In specific terms, for the flow of an electric current through a conducting medium, equation (10.1) takes the form:

$$I = \frac{\sigma a V}{l} \quad (10.2)$$

where I is the current (A), σ is the conductivity of the medium (S m^{-1}; i.e. the field characteristic), a is the cross-sectional area of the medium (m^2), l is the length of the medium (m), and V is the applied potential difference, or voltage (V). The group ($\sigma a/l$) is termed the 'conductance', denoted by G and measured in siemens, thus:

$$I = GV \quad (10.3)$$

The reciprocal of conductance is referred to as the 'resistance', denoted by R, and measured in ohms. Hence

$$I = V/R \quad (10.4)$$

Equation (10.4) is the familiar *Ohm's law*, which defines a linear relationship between voltage and current in a conducting medium. If the resistance R varies with the magnitude of the voltage, or the current, then the resistance is non-linear. Rectifiers constitute one particular class of non-linear resistors.

Comparing equations (10.4) and (10.2) gives:

$$R = l/(\sigma a) \quad (10.5)$$

It is more usual, however, to quote the 'resistivity' as opposed to the conductivity, and resistance is generally written as:

$$R = \rho l/a \quad (10.6)$$

where ρ is the resistivity of the conductor (in Ω-m).

The resistance of all pure metals is temperature dependent, increasing linearly for moderate increases in temperature. Other materials, including carbon and many insulators, exhibit a decreasing resistance for an increase in temperature.

10.2.2 Simple resistive circuits

The effective total resistance of the series arrangement shown in *Figure 10.1* is the algebraic sum of all the resistances in series, i.e.

$$R_t = R_1 + R_2 + R_3 \quad (10.7)$$

where R_t is the total resistance of the circuit.

In *Figure 10.2*, the resistors are connected in parallel and each resistor is subject to the same applied voltage. The effective total resistance obeys an inverse summation law, i.e.

$$\frac{1}{R_t} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} \quad (10.8)$$

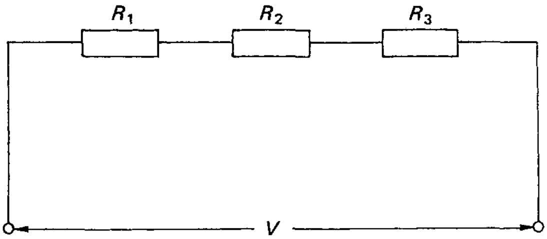


Figure 10.1 Resistors in series

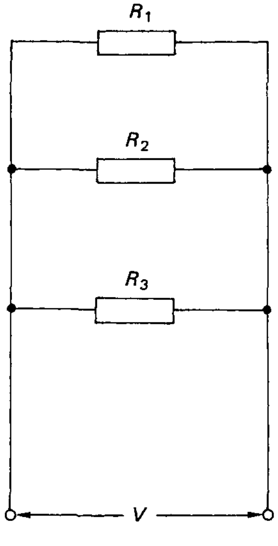


Figure 10.2 Resistors in parallel

Using equations (10.7) and (10.8), many series parallel resistive circuits can be reduced to a single effective resistance. Back calculation can then yield the current and potential differences across all the resistive elements which make up the circuit.

10.2.3 Electromotive force and potential difference

In a metallic conductor which has a potential difference applied across opposite ends, free electrons are attracted to the more positive end of the conductor. It is this drift of electrons which constitutes the electric current and the effect is simply Nature's attempt to redress an energy imbalance. Although the negatively charged electrons actually drift towards the positive end of the conductor, traditional convention gives the direction of the current flow from positive to negative. There is nothing really at issue here since it is only a simple sign convention which was adopted long before the true nature of the atom and its associated electrons was postulated.

A current of 1 A is associated with the passage of 6.24×10^{18} electrons across any cross-section of the conductor per second. The quantity of charge is the coulomb Q:

$$Q = It \qquad (10.9)$$

where 1 C of charge is passed when a current of 1 A flows for a period of 1 s.

The electromotive force (e.m.f.) is that which tends to produce an electric current in a circuit. Since e.m.f. is also measured in volts, the distinction between electromotive force and potential difference might at first appear rather subtle. However, e.m.f. is associated with the energy source. Potential difference is simply the product of current and resistance across any resistive element in a circuit, irrespective of the energy source.

For circuit elements other than purely resistive ones, the potential difference across the element becomes a time-dependent function.

10.2.4 Power and energy

Power is the rate at which energy is expended or supplied. The potential difference across any two points in a circuit is defined as the work done in moving unit charge from a lower to a higher potential. Thus the work done in moving Q coulomb of charge across a constant potential difference of V volt is:

$$W = QV \qquad (10.10)$$

Therefore

$$\text{Power} = \frac{dW}{dt} = \frac{dQ}{dt}V$$

From equation (10.9),

$$(dQ/dt) = I$$

Thus

$$\text{Power} = IV \qquad (10.11)$$

Using Ohm's law, the power dissipated across a simple resistive circuit element is:

$$\text{Power} = IV = I(IR) = I^2R \qquad (10.12)$$

10.2.5 Network theorems

A network consists of a number of electrical elements connected up in a circuit. If there is no source of e.m.f. in the circuit, it is said to be 'passive'. When the network contains one, or more, sources of e.m.f., it is said to be 'active'.

Many practical resistive networks cannot be simplified to equivalent series and parallel arrangements. The unbalanced Wheatstone bridge provides one classic and frequently encountered example. For this reason a number of well-established theorems have been developed for the analysis of such complex resistive networks (see Hughes,[3] Bell and Whitehead,[4] or Bell[5]). The theorems are listed below.

(1) *Kirchhoff's first law*—The algebraic sum of the currents entering (positive) and leaving (negative) a junction is zero.
(2) *Kirchhoff's second law*—The algebraic sum of potential differences and e.m.f.s around any closed circuit is zero.
(3) *Superposition theorem*—In a linear resistive network containing more than one source of e.m.f., the resultant current in any branch is the algebraic sum of the currents that would be produced by each e.m.f. acting on its own while the other e.m.f.s are replaced with their respective internal resistances.
(4) *Thevenin's theorem*—The current through a resistor R connected across any two points in an active network is obtained by dividing the potential difference between the two points, with R disconnected, by $(R + r)$, where r is the resistance of the network between the two connection points with R disconnected and each e.m.f. replaced with its equivalent internal resistance.
(5) *Norton's theorem*—Any active network can be replaced at any pair of terminals by an equivalent current source in parallel with an equivalent resistance.

Although perhaps long-winded in definition, Thevenin's theorem has great practical application to complex resistive networks. An alternative statement of Thevenin's theorem is: 'Any active network can be replaced at any pair of terminals by an equivalent e.m.f. in series with an equivalent resistance'. The more concise version of Thevenin's theorem is perhaps a little more indicative of its power in application.

It might be apparent that Norton's theorem is complementary to Thevenin's theorem and both can be equally well used in the analysis of resistive networks.

Other useful network analysis techniques include 'mesh analysis', which incorporates Kirchhoff's first law, and 'nodal analysis' which is based on Kirchhoff's second law. Mesh and nodal analysis are also essentially complementary techniques.

A proficiency in the use of the network theorems can only be obtained through experience in their application. The books by Hughes[3] and Bell and Whitehead[4] provide a wealth of worked and tutorial examples with answers for practice.

10.2.6 Double-subscript notation

To avoid ambiguity in the direction of current, e.m.f. or potential difference, a double-subscript notation has been adopted. *Figure 10.3* shows a source of e.m.f. which is acting from D to A. The e.m.f. is therefore E_{da}. The current flows from A to B, by traditional convention, and is designated I_{ab}. From this simple circuit it is apparent that $I_{ab} = I_{bc} = I_{cd} = I_{da}$.

The potential difference across the load R is denoted by V_{bc} to indicate that the potential at B is more positive than that at C. If arrow heads are used to indicate the potential difference, they should point towards the more positive potential.

10.2.7 Electrostatic systems

Electrostatic systems are quantified by the physical behaviour of the somewhat intangible concept of 'charge'. Fortunately,

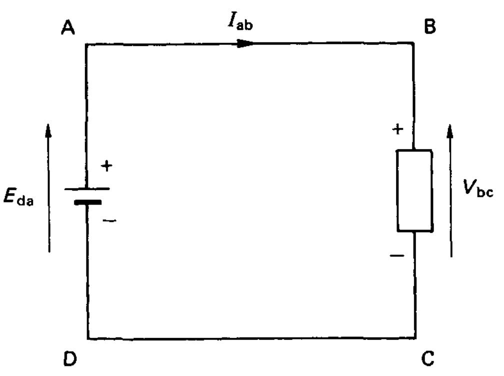


Figure 10.3 Double-subscript notation

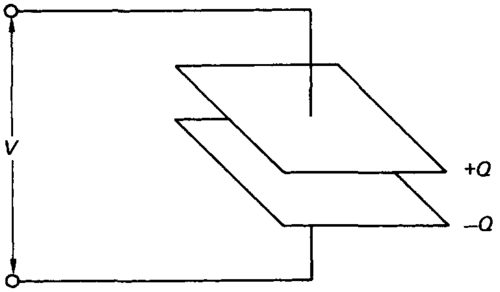


Figure 10.4 Electrostatic system

the unified field approach lends itself well to the quantification of electrostatic systems.

Figure 10.4 shows two parallel, conducting metal plates separated by an evacuated space. A potential difference is applied across the plates such that they become charged at equal magnitude, but opposite sign. For the electrostatic system, equation (10.1) is written:

$$Q = \frac{\varepsilon_0 a V}{l} \tag{10.13}$$

where Q is the total charge (C), ε_0 is the permittivity of free space in (F m^{-1}; i.e. the field characteristic), a is the cross-sectional area of the plates, l is the distance separating the plates, and V is the applied potential difference.

The term ($\varepsilon_0 a/l$) is called the 'capacitance' of the system. It is usually denoted by C, and is measured in units of farads. Thus

$$Q = CV \tag{10.14}$$

Since the Farad is an unwieldy large number, it is more common to use the microfarad (μF) or the picofarad (pF) as the unit of measurement. (N.B. 1 μF = 10^{-6} F; 1 pF = 10^{-12} F.)

If the plates are separated by an insulating medium other than free space, then these so-called 'dielectric media' have a different value of permittivity. The actual permittivity is related to the permittivity of free space by the relative permittivity of the dielectric, i.e.

$$\varepsilon = \varepsilon_0 \cdot \varepsilon_r \tag{10.15}$$

where ε_r is the relative permittivity of the dielectric.

Table 10.1 The permittivity of some typical dielectric materials

Material	*Relative permittivity*
Air	1
Paper	2–2.5
Porcelain	6–7
Mica	3–7

The permittivity of free space ε_0 is numerically equal to $(1/36\pi) \times 10^{-9}$. The relative permittivities of some of the more common dielectric materials are listed in *Table 10.1*.

10.2.8 Simple capacitive circuits

Figure 10.5 shows three capacitors connected in a simple parallel arrangement. The equivalent total capacitance is given as the algebraic sum of all the capacitances in the circuit, i.e.

$$C = C_1 + C_2 + C_3 \tag{10.16}$$

where C is the total capacitance.

For the series capacitance arrangement shown in *Figure 10.6*, the total equivalent capacitance is related through the inverse summation:

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} \tag{10.17}$$

Equations (10.16) and (10.17) can be used to reduce series and parallel capacitor circuits to a single equivalent capacitor.

Composite capacitors, involving different dielectric media, may also be treated in the same manner as a series capacitor arrangement.

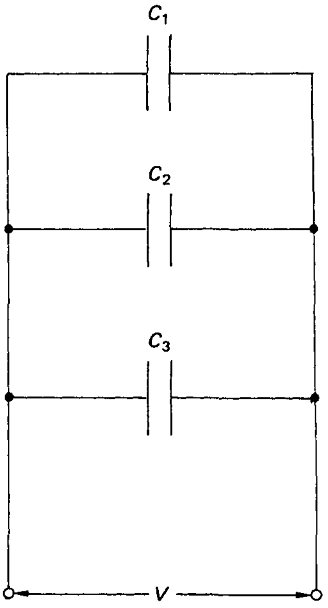


Figure 10.5 Capacitors in parallel

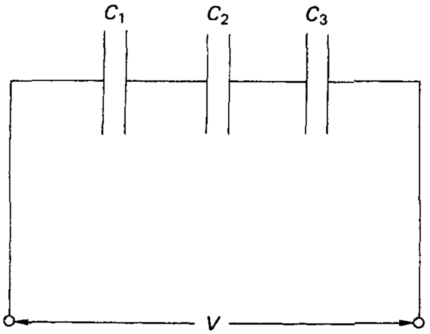


Figure 10.6 Capacitors in series

10.2.9 Charging a capacitor

Figure 10.7 shows a parallel plate capacitor which is connected in series with a resistor to a source of e.m.f., say a battery, through a switch. Initially, the capacitor is uncharged before the switch is closed. When the switch is closed a charging current will flow until such time that the potential difference across the capacitor is equal to the e.m.f. available from the source. The charging process consists of taking electrons from A and transferring them through the external wiring to plate B. The energy required to do this is derived from the battery. The build up of electrons from the negative terminal of the battery to plate B of the capacitor induces a dielectric flux between the plates and a balancing positive charge is developed on plate A. As long as the dielectric flux is changing, a current will flow externally. Eventually a state of equilibrium will be reached. Note that no electrons can pass through the dielectric because it is an insulator.

The instantaneous current during charging is:

$i = \mathrm{d}Q/\mathrm{d}t$

From equation (10.14), this may be rewritten for a capacitor as:

$$i = \mathrm{d}Q/\mathrm{d}t = C(\mathrm{d}v/\mathrm{d}t) \qquad (10.18)$$

where v is the instantaneous voltage.

The instantaneous power is therefore:

$p = iv = Cv(\mathrm{d}v/\mathrm{d}t)$

The energy supplied over the time period $\mathrm{d}t$ is:

$Cv(\mathrm{d}v/\mathrm{d}t)\mathrm{d}t = Cv\mathrm{d}v$

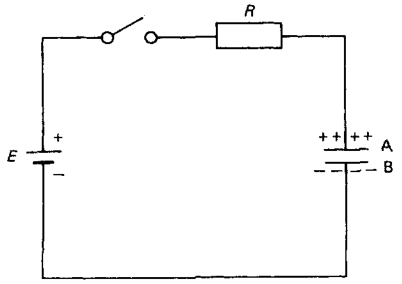


Figure 10.7 Charging a capacitor

Hence the total energy supplied is:

$$\int_0^V Cv\mathrm{d}v = \frac{1}{2}CV^2 \qquad (10.19)$$

10.2.10 Types of capacitor

10.2.10.1 Air capacitors

These usually consist of one set of fixed plates and another set of moveable plates. The area between the capacitor plates is therefore variable and so also (from equation (10.13)) is the capacitance.

10.2.10.2 Paper dielectric capacitors

These consist of metal foil interleaved with wax, or oil impregnated paper and rolled into a compact cylinder form.

10.2.10.3 Mica dielectric capacitors

These consist of alternate layers of mica and metal foil, tightly clamped together. They are used mainly for high frequency applications.

10.2.10.4 Ceramic capacitors

The electrodes are formed by a metallic coating, usually silver, deposited on the opposite faces of a thin ceramic disc.

10.2.10.5 Polycarbonate capacitors

Polycarbonate is a plastic insulating material which is produced in a range of thicknesses, down to about 2 μm. The polycarbonate is bonded with aluminium foil and rolled to form the capacitor element.

10.2.10.6 Electrolytic capacitors

These generally consist of two aluminium foils, one with an oxide film and the other without. The foils are usually interleaved with paper which is saturated with a suitable electrolyte. Electrolytic capacitors have the advantage of having a large capacitance for a relatively small physical volume. They are used extensively for smoothing out the ripple voltage from rectified a.c. power supplies. They are only suitable, however, for use in circuits where the applied voltage across the capacitor can never reverse its direction.

10.2.11 Dielectric strength

If the potential difference across opposite faces of a dielectric material is increased above a particular value, then the material breaks down. The failure of the material takes the form of a small puncture, which renders the material useless as an insulator. The potential gradient necessary to cause breakdown is normally expressed in kilovolts/millimetre (kV mm^{-1}) and is termed the 'dielectric strength'. The dielectric strength of a given material decreases with increases in the thickness. *Table 10.2* gives approximate values for some of the more common dielectric materials.

10.2.12 Electromagnetic systems

The magnetic field can be defined as the space in which a magnetic effect can be detected, or observed. The standard school physics experiment involves sprinkling iron filings on a sheet of paper under which a bar magnet is placed. On tapping

Table 10.2 The dielectric strength of some common insulators

Material	*Thickness* (mm)	*Dielectric strength* (kV mm^{-1})
Air	0.2	5.75
	0.6	4.92
	1.0	4.36
	10.0	2.98
Mica	0.01	200
	0.10	115
	1.00	61
Waxed paper	0.10	40–60

the paper lightly, the filings arrange themselves along closed loops of equal potential and are clearly visible. An obvious magnetic field is also observable around a straight length of conductor carrying a current. In particular, the exact same magnetic field as that produced by the bar magnet is observed when the current carrying conductor is formed into a helical type coil. The equipotential loops describe the path of the magnetic flux ϕ and, although the flux lines have no physical meaning, they provide a convenient vehicle to quantify various magnetic effects.

The direction of the magnetic flux is governed by the so-called 'right-hand screw rule'. This states that the direction of the magnetic field produced by a current corresponds to the direction given by turning a right-hand screw thread. The direction of the current corresponds to the translational movement of the screw.

10.2.13 Magnetic field of a toroid

Figure 10.8 shows a toroidal coil, of N turns which is wound round an annular former. A resultant magnetic flux (shown as broken lines) is generated when the coil carries a current. For the magnetic field, equation (10.1) takes the general form:

$$\phi = \frac{\mu a F}{l} \quad (10.20)$$

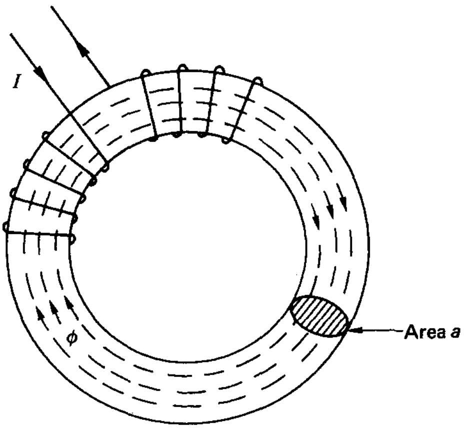


Figure 10.8 Toroid

where ϕ is the magnetic flux (W), μ is the permeability of the medium (H m^{-1}), a is the cross-sectional area of the flux path in the toroid, l is the length of the flux path, and F is the magnetic potential difference, or magnetomotive force (A).

The magnetomotive force (m.m.f.) is equal to the product of the number of turns on the coil and the current carried:

$$F = IN \quad (10.21)$$

Note that the m.m.f. is descriptively expressed in 'ampere-turns' (A-t). Since the number of turns is already a dimensionless quantity, the accepted unit of magnetomotive force is the ampere (A).

The term ($\mu a/l$) is called the 'permeance' and the inverse of permeance is the 'reluctance' S. Thus equation (10.20) may be rewritten as:

$$\phi = F/S \quad (10.22)$$

Equation (10.22) represents an electromagnetic version of Ohm's law.

Alternatively, equation (10.20) can be expressed as:

$$\frac{\phi}{a} = \mu \frac{F}{l}$$

or

$$B = \mu H \quad (10.23)$$

where $B = \phi/a$ is the magnetic flux density (W m^{-2}, or T) and $H = F/l$ is the magnetic intensity (A m^{-1}).

10.2.14 Permeability

The permeability of free space μ_0 is numerically equal to $4\pi \times 10^{-7}$. The absolute permeability of other materials is related to the permeability of free space by the relative permeability:

$$\mu = \mu_0 \mu_r \quad (10.24)$$

For air and other non-magnetic materials, the absolute permeability is the same constant. For magnetic materials, absolute permeability is not a fixed constant but varies non-linearly with the flux density. The non-linear variation of permeability is conveniently displayed as a functional plot of magnetic flux density B against magnetic intensity H. *Figure 10.9* illustrates a number of B vs. H curves for some common materials. From equation (10.23) it is apparent that the absolute permeability is given by the slope of a tangent to the B vs. H curve at any particular value.

Also shown in *Figure 10.9* is the B vs. H curve for air, the only straight-line relationship in the diagram. It is apparent that, for an applied magnetic intensity, the magnetic flux developed in a coil with a ferrous core is many times greater than that through a similar coil with an air core. In most practical systems, therefore, a ferrous core is normally used since it greatly facilitates the establishment of a magnetic flux.

10.2.15 Faraday's law

Perhaps the most fundamental law of electromagnetic systems, Faraday's law states that the e.m.f. induced in a magnetic circuit is equal to the rate of change of flux linkages in the circuit. In mathematical terms, Faraday's law is given by:

$$e = N(d\phi/dt) \quad (10.25)$$

where e is the instantaneous induced e.m.f.

Equation (10.25) forms the basis of all electrical power generating machines and is a statement of the fact that an

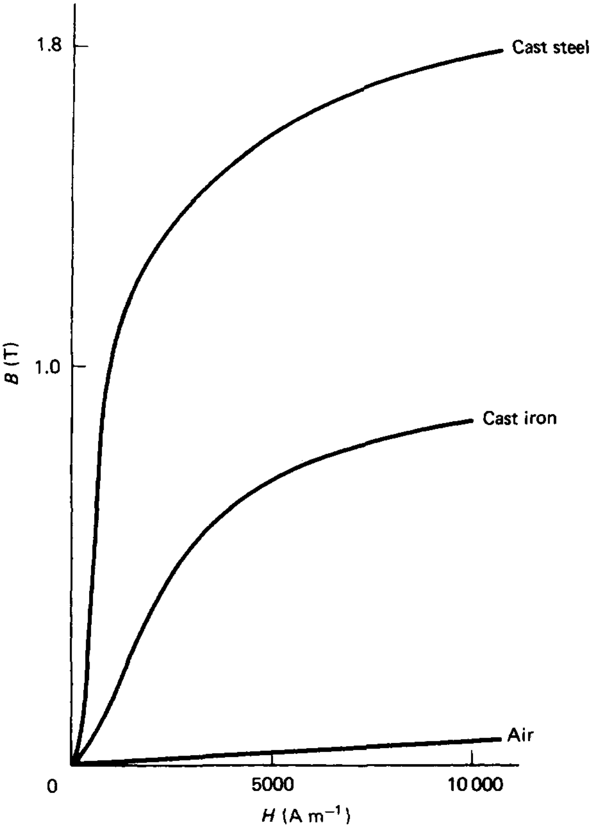


Figure 10.9 Plots of magnetic flux density B vs. magnetic intensity H for some common materials

electric current can be produced by the movement of magnetic flux relative to a coil. In all rotating electrical generators, it is actually the coil which is moved relative to the magnetic field. The net result, however, is exactly the same.

The direction of the induced e.m.f. is always such that it tends to set up a current to oppose the motion, or the change of magnetic flux which is responsible for inducing the e.m.f. This is essentially a statement of Lenz's law. In many texts, therefore, the right-hand side of equation (10.25) is often shown as a negative quantity.

The motion, or change of flux is associated with the application of a mechanical force which ultimately provides the torque required to drive the electric generator. *Figure 10.10* shows a single conductor of length l metres, carrying an induced current I and lying perpendicular to a magnetic field of flux density B tesla. The force applied causes the conductor to move through a distance dx metres. The mechanical work done is, therefore, Fdx. The electrical energy produced is given as the product of the power developed and the time duration, i.e. $eIdt$. For no external losses, the mechanical work done is converted into electrical energy. Thus

$$eIdt = Fdx \qquad (10.26)$$

Using Faraday's law (equation (10.25)) the induced e.m.f. is equal to the rate of change of flux linkage. For a single conductor, $N = 1$ and thus

$$e = (Bldx)/dt$$

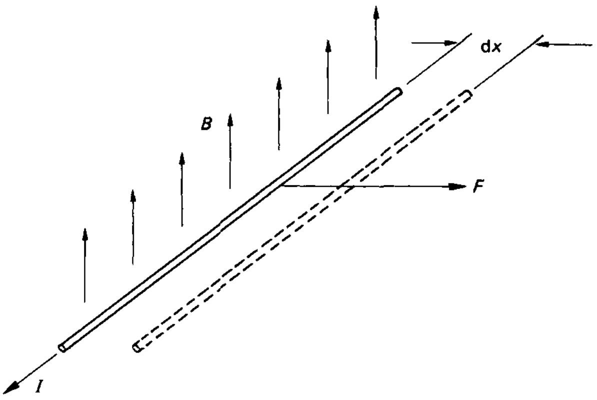


Figure 10.10 Generation of e.m.f.

Therefore

$$(Bldx/dt)Idt = Fdx$$

i.e.

$$F = BlI \qquad (10.27)$$

Equation (10.27) relates the applied force to the corresponding current generated in a conductor moving through a magnetic field. The equation applies equally to an electric generator, or conversely to a motor in which case the electrical power supplied is converted into a mechanical torque via the electromagnetic effect.

10.2.16 Self-induced e.m.f.

If a current flows through a coil, a magnetic flux links with that coil. If, in addition, the current is a time varying quantity, then there will be a rate of change of flux linkages associated with the circuit. The e.m.f. generated will oppose the change in flux linkages.

When dealing with electric circuits it is convenient if the voltage across individual elements can be related to the current flowing through them. *Figure 10.11* shows a simple circuit comprising a coil having N turns and resistance R, connected in series with a time varying voltage.

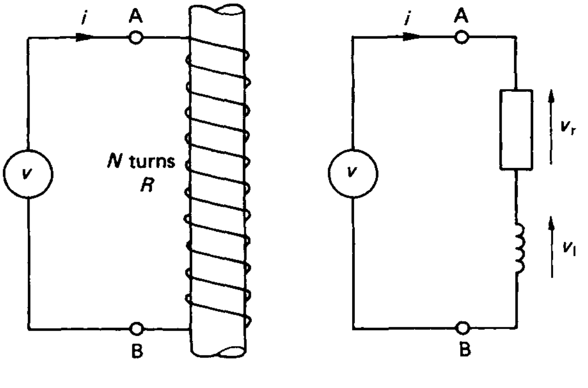


Figure 10.11 Self-induced e.m.f.

The voltage drop across the terminals A and B can be split into two components. First there is the voltage drop purely due to the resistance of the coiled element. Secondly there is a voltage drop which is a consequence of the self-induced e.m.f. generated through the electromagnetic effect of the coil. Thus

$$v = v_r + v_l$$
$$= iR + N\frac{d\phi}{dt} \quad (10.28)$$

From equations (10.20) and (10.21)

$$\phi = \frac{\mu aF}{l} = \frac{\mu aiN}{l}$$

therefore

$$v = iR + N\frac{d}{dt}\left(\frac{\mu aiN}{l}\right)$$
$$= iR + N^2\left(\frac{\mu a}{l}\right)\frac{di}{dt} \quad (10.29)$$

The group $N^2(\mu a/l)$ is called the 'self-inductance' of the coil and is denoted by L. The unit of self-inductance is the henry. Therefore,

$$v = iR + L\frac{di}{dt} \quad (10.30)$$

By comparing equations (10.28) and (10.30) it is apparent that

$$L\frac{di}{dt} = N\frac{d\phi}{dt}$$

Integration then gives

$$L = N\phi/i \quad (10.31)$$

The nature of the self-induced e.m.f., i.e. (Ldi/dt), is such that it will oppose the flow of current when the current is increasing. When the current is decreasing the self-induced e.m.f. will reverse direction and attempt to prevent the current from decreasing.

10.2.17 Energy stored in an inductor

Instantaneous power = vi

$$\text{Energy stored} = W = \int_0^t vi\,dt$$
$$= \int_0^t L\frac{di}{dt}\,i\,dt$$
$$= L\int_0^I i\,di = \frac{1}{2}LI^2 \quad (10.32)$$

10.2.18 Mutual inductance

Two coils possess mutual inductance if a current in one of the coils produces a magnetic flux which links the other coil. *Figure 10.12* shows two such coils sharing a common magnetic flux path in the form of a toroid.

The mutual inductance (in henry) between the two coils is:

$$M = \frac{N_2\phi}{I_1} \quad (10.33)$$

where N_2 is the number of turns on coil 2, I_1 is the current through coil 1, and ϕ is the magnetic flux linking coils 1 and 2.

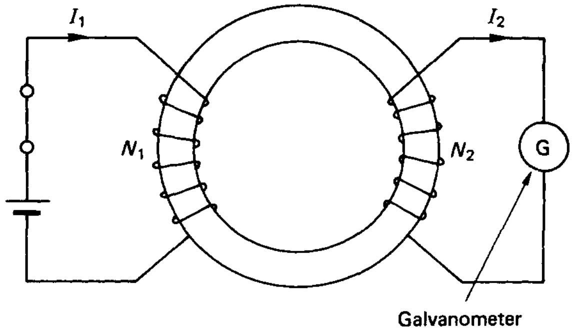


Figure 10.12 Mutual inductance

The mutual-inductance effect finds wide application both to electrical transformers and to rotating electrical machines.

10.2.19 Hysteresis in magnetic circuits

Hysteresis can be described with reference to a toroidal coil wound on an iron core (see *Figure 10.8*). The current applied to the coil can be imagined to be taken through a cyclic process where it is increased from 0 to $+I$ amperes, back through 0 to $-I$ amperes and again back through 0 to $+I$ amperes. Measurement of the flux density in the core, as the current varies, results in a B vs. H curve as depicted in *Figure 10.13*.

The behaviour of the B vs. H relationship is termed a 'hysteresis loop'. This behaviour is typical for ferrous cores and is an illustration of the fact that all of the electrical energy supplied to magnetise an iron core is not returned when the coil current is reduced to zero. The loss of energy is called 'hysteresis loss' and it is manifested as heat in the iron core. It can be shown that the hysteresis loss is directly proportional to the area enclosed by the hysteresis loop. It is to obvious

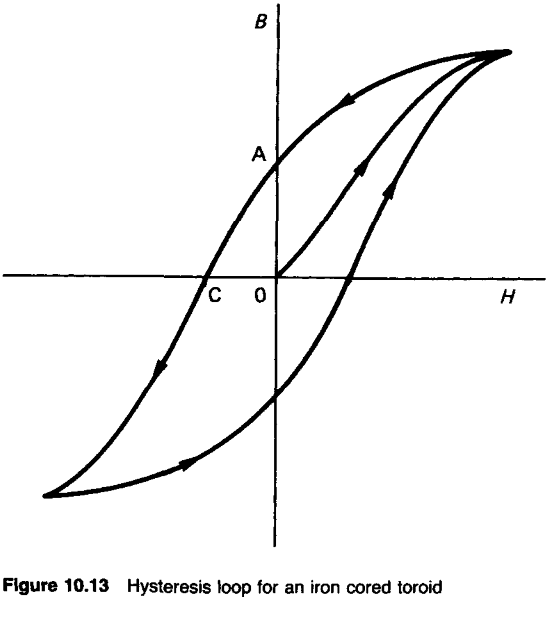


Figure 10.13 Hysteresis loop for an iron cored toroid

advantage, therefore, that any magnetic system which is subject to a cyclic variation of flux density should incorporate a magnetic core in which the hysteresis loop is minimally small. In practical applications a low-loss core of silicon iron, or a nickel–iron alloy such as Mumetal or Permalloy, is normally used. Some strain-relieving heat treatment process is also usually involved, after cold working, to restore the original magnetic condition of the material.

Hysteresis is characterised by two parameters: the 'remanent flux density' (or 'remanence') and the 'coercive force'. The remanent flux density is the flux density which remains in the core when the mangetic intensity, i.e. the coil current, has been reduced to zero. The remanent flux density is represented by line 0A in *Figure 10.13*. The coercive force is the magnetic intensity required to reduce the remanent flux density to zero and is represented by line 0C in *Figure 10.13*.

10.2.20 Eddy current loss

Faraday's law (equation (10.25)) shows that a time varying magnetic flux will induce an e.m.f. in a coil. If the ends of the coil are connected and form a closed circuit, then the induced voltage will circulate a current round the closed loop. Let us now consider an iron core, in which a time varying magnetic flux exists. Since iron is a conductor, then there will be a multitude of arbitrary closed paths within the iron matrix. These closed paths constitute effective conduction routes and the varying magnetic flux will generate a flow of current round them. The currents are called 'Eddy currents', and because of the ohmic resistance of the core, the end result is an energy loss as the eddy currents are dissipated as heat.

Eddy current losses can be greatly reduced by building the iron core in the form of laminations which are insulated from one another. The laminated assembly confines the path lengths for the eddy currents to each respective lamination. The cross-sectional area of the eddy current path is also reduced and the eddy current loss is approximately proportional to the square of the thickness of the laminations. A practical minimum thickness for any lamination is about 0.4 mm. Increasing manufacturing costs could not justify the use of much thinner laminations.

The eddy current phenomenon, while basically a detrimental effect, can be utilised to good purpose in the form of 'induction heating'. In this type of heater, a metallic work piece is enclosed within a coil which is powered with an a.c. supply. The induced eddy currents are dissipated as heat in the work piece. Induction heating is basically a surface heating technique where the 'depth' of heating is dependent on the depth of eddy current activity. This, in turn, is dependent on the frequency of the a.c. supplied to the coil. A high frequency supply produces a relatively low depth of heat penetration and this is used extensively in surface, or case hardening heat treatment processes.

10.2.21 Kirchhoff's laws and the magnetic circuit

Figure 10.14 shows a magnetic circuit in which a magnetising coil is wound on one of the limbs and another limb incorporates the usual features of an 'air gap'.

Using the analogy between the magnetic and the conduction circuits, the magnetic circuit can be represented in terms of an energy source, or m.m.f., and each limb of the magnetic circuit is written in terms of the appropriate reluctance S. This is illustrated in *Figure 10.15*.

Given all the relevant dimensions and material properties, the problem resolves to one of calculating the current required to establish a prescribed magnetic flux density in the air gap.

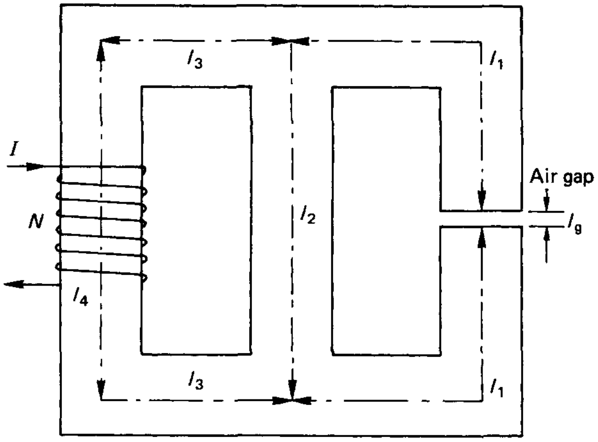


Figure 10.14 Magnetic circuit

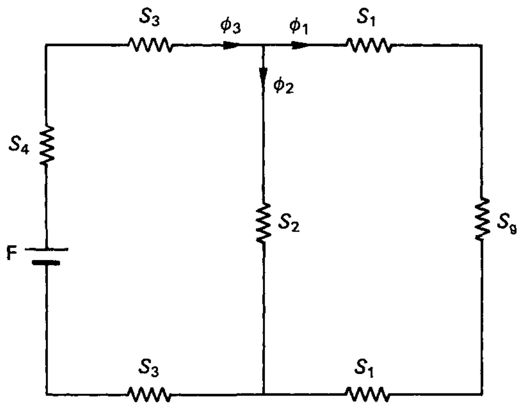


Figure 10.15 Representation of the magnetic circuit shown in *Figure 10.14*

The solution invokes the use of Kirchhoff's laws as they apply to magnetic circuits.

First law— At any instant in time, the sum of the fluxes flowing into a node is equal to the sum of the fluxes flowing out.

Second law—Around any closed magnetic circuit the total magnetomotive force is equal to the sum of all the m.m.f.s round the circuit.

Manipulation of equations (10.20) to (10.24) then yields the required solution. The self-inductance of the coil, if required, may be calculated from equation (10.31), or from the definition:

$$L = N^2(\mu a/l) = N^2/S \qquad (10.34)$$

It has already been shown that the lowest permeability is that of air and that the m.m.f. required to produce a flux density in air is many times greater than that required to produce the same flux density in a ferrous material. It may reasonably be questioned, therefore, why air gaps are used at

all in iron cored magnetic circuits. The only function of the air gap is to provide a measure of linearity to the magnetic system such that the inductance remains reasonably constant over a range of operating currents.

10.2.22 Alternating quantities

If an electrical quantity varies with time, but does not change its polarity, it is said to be a 'direct current' (d.c.) quantity. If the quantity alternates between a positive and a negative polarity, then it is classified as an 'alternating current' (a.c.) quantity.

The period T is the time interval over which one complete cycle of the alternating quantity varies. The inverse of the period is the frequency, f (in cycles per second, or Hertz). The circular frequency ω in radians per second is also commonly used. (N.B. One cycle/second corresponds to 2π rad s^{-1}.)

Instantaneous values of the quantities encountered in electrical systems are usually denoted by lower-case letters. Since the instantaneous values are difficult to measure and quantify, a.c. quantities are usually expressed as 'root mean square' (r.m.s.) values. For a periodically varying a.c. quantity, the r.m.s. value is given by:

$$\text{r.m.s.} = \left[1/t \int_0^t (\text{Quantity})^2 dt\right]^{1/2} \qquad (10.35)$$

If in a periodically varying a.c. quantity, the positive and negative half cycles are identical in all but sign, then the average value of the quantity is zero. The average value of the half cycle is sometimes useful, however, particularly when dealing with rectified signals.

Many electrical quantities vary in a sinusoidal manner and it can easily be shown that the r.m.s. value is simply related to the maximum value by:

$$\text{r.m.s.} = \text{max.}/(\sqrt{2}) = 0.707 \text{ max.} \qquad (10.36)$$

10.2.23 Relationship between voltage and current in resistive, inductive and capacitive elements

For a simple resistive element, current is directly proportional to voltage. The current waveform will therefore be essentially the same shape as the voltage waveform.

For an inductive coil with negligible resistance, the relationship between voltage and current is given by equation (10.30), i.e.

$$v = L\frac{di}{dt}$$

Thus

$$i = \frac{1}{L}\int v dt \qquad (10.37)$$

If, for example, the voltage is an a.c. square wave, then the current will take the shape of a 'sawtooth' waveform (see *Figure 10.16*).

The relationship between voltage and current for a capacitive element is given by equation (10.18), i.e.

$$i = C\frac{dv}{dt}$$

For the capacitive element it can be seen that a current will flow only when the voltage is changing. No current can flow if the voltage is constant since (dv/dt) will then be equal to zero. The capacitor then, will block any steady d.c. input and indeed is sometimes used for this express purpose.

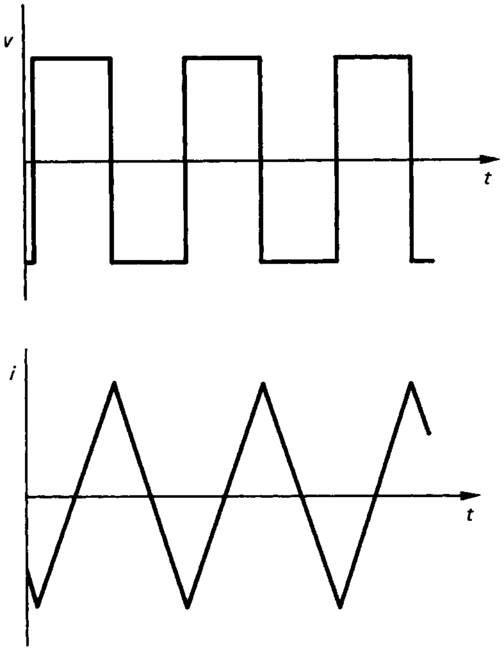


Figure 10.16 Voltage–current relationship for an inductive element

10.2.24 Resistive–inductance and resistive–capacitance circuits under transient switching conditions

Circuits involving a single resistor (R), capacitor (C) or inductance (L) are rare. It is more usual to find circuits involving some combination of these elements in both d.c. and a.c. applications. *Figure 10.17* illustrates two simple RL and RC circuits.

10.2.24.1 RL circuit

With the switch open there is no flow of current in the circuit. At the instant of switching, the current will rise and eventually reach a steady-state value of V_s/R. The transient period is governed by equation (10.30) which represents a first order, ordinary differential equation in i. The solution of equation (10.30) involves the technique of separating the variables to allow integration. The general solution is:

$$i = I[1 - \exp(-Rt/L)] \qquad (10.38)$$

where I is the final steady-state current in the circuit (see *Figure 10.18*).

Equation (10.38) shows that the current growth in the circuit will rise exponentially to reach a steady-state value as time t increases.

It may also be shown that:

$$v = L\frac{di}{dt} = V_s \exp(-Rt/L) \qquad (10.39)$$

If the initial rate of increase in current is maintained, then the steady-state current is achieved in a time T (seconds). From *Figure 10.18*:

$$\left[\frac{di}{dt}\right]_{t=0} = \frac{I}{T}$$

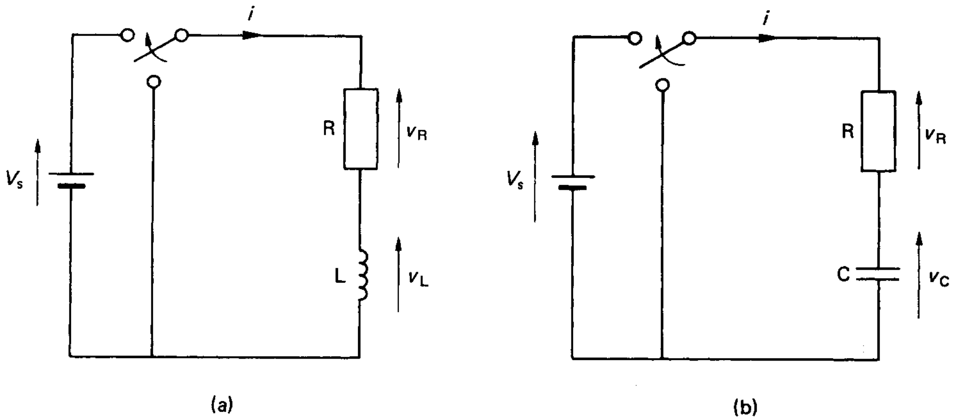


Figure 10.17 Simple RL (a) and RC (b) circuits under transient switching conditions

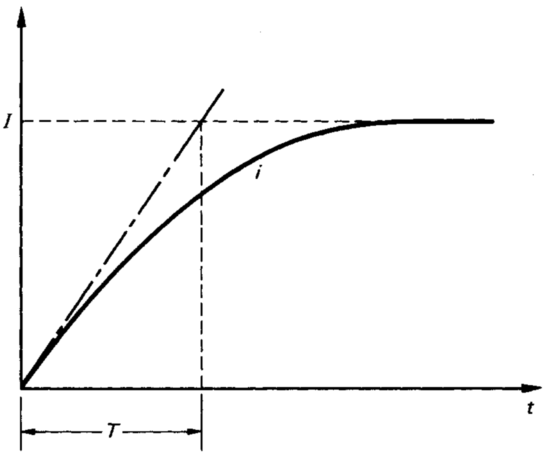


Figure 10.18 Transient current growth in an inductive element

At the instant of switching, $v_L = V_s$. Hence at $t = 0$

$$v_L = V_s = LI/T$$

Therefore

$$T = LI/V_s = L/R \qquad (10.40)$$

Equation (10.40) defines the 'time constant' for the RL circuit. The actual value of the current in the circuit when $t = T$ is $0.632 \times I$.

10.2.24.2 RC circuit

In *Figure 10.17(b)*, with the switch open there is zero potential difference across the capacitor. On closing the switch the voltage across the capacitor will rise in an asymptotic manner, reaching a steady-state value of V_s. From Kirchhoff's second law:

$$V_s = iR + v_c \qquad (10.41)$$

where v_c is the instantaneous voltage across the capacitor.

From equation (10.18) we can write:

$$V_s = RC\frac{dv_c}{dt} + v_c \qquad (10.42)$$

Equation (10.42) shows that the instantaneous voltage across the capacitor also conforms to a first-order system. The solution of (10.42) gives

$$v_c = V_s[1 - \exp(-t/RC)] \qquad (10.43)$$

Considering the initial rate of increase of voltage, it can be shown that the appropriate time constant for the simple RC circuit is:

$$T = RC \qquad (10.44)$$

Both the simple RL and RC circuits are first-order systems with a generalised form of transient behaviour.

In circuits containing both inductive and capacitive elements, the transient behaviour is governed by a second order ordinary differential equation. The transient behaviour of these circuits, however, is less important than their response to sinusoidally varying inputs.

An interesting additional feature of the RC circuit is its ability to function as an integrating, or a differentiating circuit. *Figure 10.19* depicts an RC integrator.

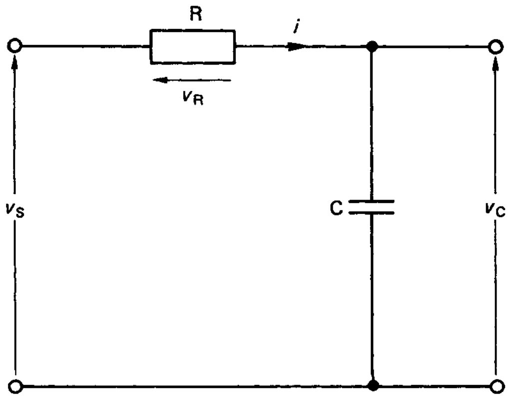


Figure 10.19 RC integrator

The components of the circuit are selected such that $v_R \gg v_c$. To a reasonable approximation $v_s = v_R$. Therefore

$$i = v_s/R$$

Since

$$v_c = 1/C \int i\mathrm{d}t$$

Then

$$v_c = 1/RC \int v_s\mathrm{d}t \qquad (10.45)$$

Equation (10.45) shows that the output voltage is proportional to the integral of the input voltage, with (1/*RC*) the constant of proportionality.

For the RC differentiator (*Figure 10.20*) the components are selected such that the voltage drop across the capacitor v_C is very much greater than that over the resistor v_R, i.e. $v_s \gg v_c$. Therefore

$$v_s = 1/C \int i\mathrm{d}t$$

Thus

$$i = C\frac{\mathrm{d}v_s}{\mathrm{d}t}$$

Taking the output across the resistor gives:

$$v_R = iR = RC\frac{\mathrm{d}v_s}{\mathrm{d}t} \qquad (10.46)$$

It is apparent that the output is proportional to the first derivative of the input.

10.2.25 Steady-state a.c.

In most practical applications in electrical engineering, the voltages and currents are sinusoidal. A simple series RLC circuit is depicted in *Figure 10.21*. Since the current is common to each of the circuit elements, it is used for reference purposes. The instantaneous current is defined as:

$$i = I_m \sin(\omega t) \qquad (10.47)$$

where I_m is the maximum value (or peak value) of the current, and ω is the angular, or circular frequency (in rad s^{-1}).

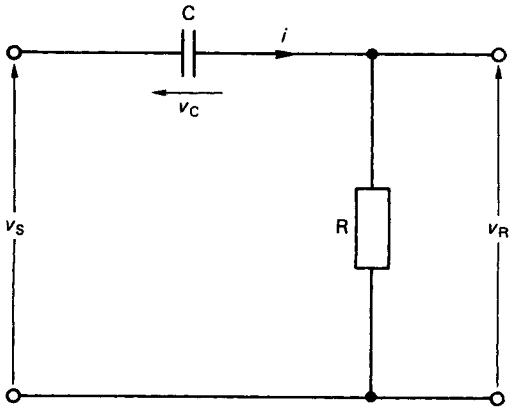


Figure 10.20 RC differentiator

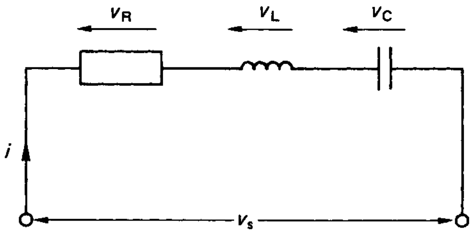


Figure 10.21 Series RLC circuit

The voltage drop across the resistor is:

$$v_R = iR = I_mR\sin(\omega t) \qquad (10.48)$$

Equation (10.48) indicates that the voltage drop across the resistor is in phase with the current. In other words, v_R reaches a positive maximum at the same instant as the current *i*.

The voltage drop across the inductor is:

$$v_L = L\frac{\mathrm{d}i}{\mathrm{d}t} = L\frac{\mathrm{d}}{\mathrm{d}t}[I_m \sin(\omega t)]$$
$$= LI_m\omega\cos(\omega t)$$
$$= \omega LI_m \sin(\omega t + 90) \qquad (10.49)$$

The relationship between current and voltage drop across the inductor is shown in *Figure 10.22*. It can be seen that there is a phase difference between the voltage drop and the current through the inductor. In fact, v_L reaches a positive maximum 'before' *i*, and v_L is said to 'lead' the current by 90°.

For the capacitor, the voltage drop is given by:

$$v_c = 1/C\int i\mathrm{d}t = 1/C\int I_m\sin(\omega t)\mathrm{d}t$$
$$= -\frac{I_m}{\omega C}\cos(\omega t)$$
$$= +\frac{I_m}{\omega C}\sin(\omega t - 90) \qquad (10.50)$$

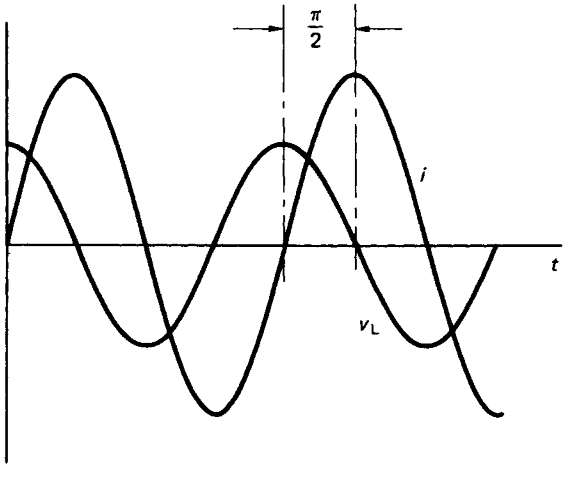


Figure 10.22 Current and voltage drop across an inductor

The voltage drop across the capacitor therefore reaches its positive maximum after that of i. In general terminology, v_c 'lags' i by 90°.

Equations (10.48) to (10.50) are all of a similar form in that they can be expressed as:

Voltage drop = Constant × Current

In equation (10.49), the constant ωL is termed the 'inductive reactance', and is denoted by X_L. In equation (10.50), the constant $(1/\omega C)$ is the 'capacitive reactance', which is denoted as X_c. Both these reactances are in units of ohms.

The total voltage drop across the three circuit elements is:

$$v = v_R + v_L + v_c$$

$$= iR + L\frac{di}{dt} + \frac{1}{C}\int i dt$$

Therefore,

$$v = I_m R \sin(\omega t) + \omega L I_m \sin(\omega t + 90°) + \frac{I_m}{\omega C}\sin(\omega t - 90) \quad (10.51)$$

While equation (10.51) defines the total instantaneous voltage drop in mathematical terms, it is a little cumbersome to deal with. To simplify the analysis, the addition of a.c. voltages is conveniently performed using a graphical technique involving 'phasors'.

10.2.26 Phasor diagrams

Any sinusoidally varying quantity can be represented as a phasor, which is a vector quantity. The length of the phasor is proportional to the magnitude of the product of the reactance and the maximum current. The direction of the phasor is determined by the phase angle and its relation to some common reference.

For the RLC circuit shown in *Figure 10.21*, the voltage drop across the inductance may be arbitrarily assumed to be greater than that across the capacitor. The total voltage drop in the circuit is then given as the phasor addition of the three individual potential difference components. This is illustrated in *Figure 10.23*.

The vector addition of the three phasors shows that the source voltage leads the current by an angle of ϕ degrees, i.e.

$$\bar{V} = V_m \sin(\omega t + \phi) \quad (10.52)$$

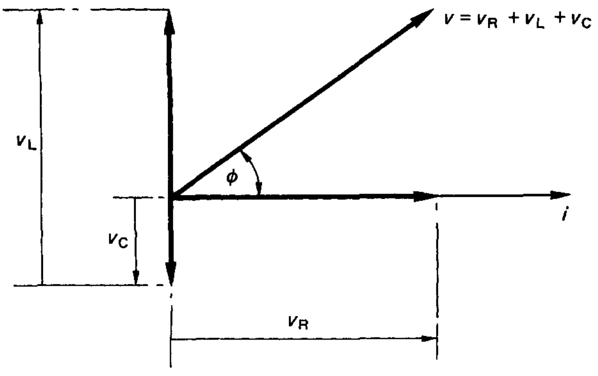


Figure 10.23 Phasor diagram for series RLC circuit

The circuit is therefore essentially inductive and, using the standard notation, the total phasor voltage is designated by a capital letter with an overscore.

10.2.27 Complex notation

Since inductive and capacitive elements in a.c. circuits involve a phase shift of +90° and −90°, respectively, the complex number notation is used extensively to manipulate phasor quantities.

The complex operator j, defined as $\sqrt{-1}$, is a unit operator which, when multiplying a phasor, shifts it by 90° in an anticlockwise direction. Thus for the series RLC circuit:

$$\bar{V}_R = \bar{I}R;\ \bar{V}_L = j\bar{I}X_L;\ \bar{V}_C = -j\bar{I}X_C$$

where $\bar{I}$ can be taken as the r.m.s. value of the current.

The voltage drop across the complete circuit can then be written as:

$$\bar{V} = \bar{I}R + j\bar{I}X_L - j\bar{I}X_C$$

$$= \bar{I}[R + j(X_L - X_C)] \quad (10.53)$$

The term within the square brackets is called the 'impedance' of the circuit and is denoted by $\bar{Z}$. Thus

$$\bar{V} = \bar{I}\bar{Z} \quad (10.54)$$

Equation (10.54) represents Ohm's law for a.c. circuits.

The phase angle between the source voltage and the current is:

$$\phi = \tan^{-1}[(X_L - X_C)/R] \quad (10.55)$$

10.2.28 The parallel resistive–inductive–capacitance circuit

A parallel RLC circuit is shown in *Figure 10.24*. The applied voltage is common to all of the circuit elements and it is therefore chosen as the reference.

Using Ohm's law, the currents through each of the circuit elements are:

$$\bar{I}_R = \bar{V}/R;\ \bar{I}_L = \bar{V}/X_L;\ \bar{I}_C = \bar{V}/X_C$$

Applying Kirchhoff's first law, the total current is the vector sum of the three currents $\bar{I}_R$, $\bar{I}_L$ and $\bar{I}_C$. The magnitude and phase of the total current may subsequently be determined

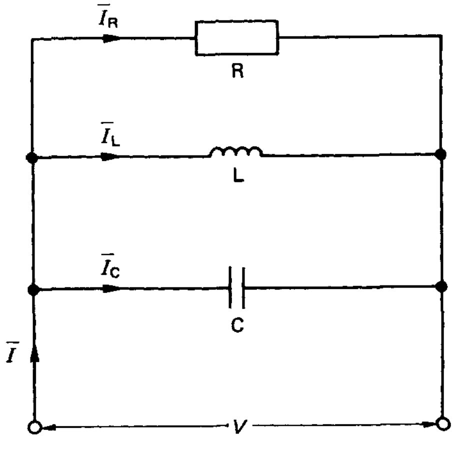


Figure 10.24 Parallel RLC circuit

from a phasor diagram, or calculated using the complex number notation. Using the latter, and noting that the current through an inductor lags the voltage while the current through a capacitor leads the voltage, it may be shown that:

$$\bar{I} = \bar{I}_R + \bar{I}_L + \bar{I}_C$$

$$= \bar{V}\left[\frac{1}{R} - j\left(\frac{1}{X_C} - \frac{1}{X_L}\right)\right] \quad (10.56)$$

and the phase angle is given by

$$\phi = \tan^{-1}\left[\frac{R(X_L - X_C)}{X_L X_C}\right] \quad (10.57)$$

10.2.29 Power and power factor in a.c. circuits

In general, for any a.c. circuit there will be a phase difference between the voltage and the current. If the circuit elements are purely resistive, the phase difference will be zero. In all other cases the phase difference will be a finite positive or negative value. Denoting the phase angle between the voltage and the current by ϕ, it may be shown (see Bell and Whitehead[4]) that the average power is:

$$P_{av} = \frac{V_m}{\sqrt{2}}\frac{I_m}{\sqrt{2}}\cos\phi$$

In terms of r.m.s. values:

$$P_{av} = VI\cos\phi \quad (10.58)$$

The term $\cos\phi$ is called the 'power factor'.

Power factor is an important parameter when dealing with electrical transformers and generators. All such machines are rated in terms of kilovolt-amperes (kV-A), which is a measure of the current-carrying capacity for a given applied voltage. The power than can be drawn depends both on the kV-A rating and the power factor of the load. *Figure 10.25* shows the relationship between kV-A, kilowatts (kW) and power factor, sometimes referred to as the 'power triangle'.

It can readily be seen that:

$$\text{kW} = \text{kV-A}\cos\phi \quad (10.59)$$

and

$$\text{kV A}_R = \text{kV-A}\sin\phi \quad (10.60)$$

where kV-A_R is the reactive power.

Thus, knowing the kV-A rating and the power factor of a number of various loads, the power requirements from a common supply may be determined.

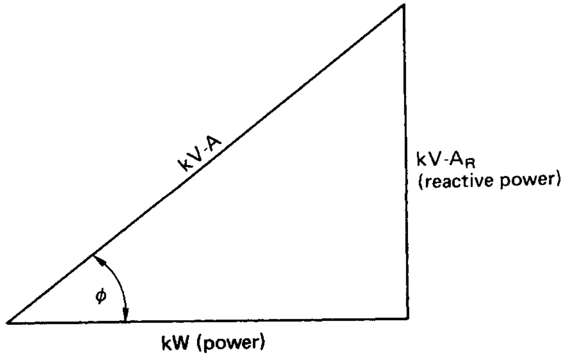


Figure 10.25 Power triangle

When quoting power factor in practical applications, it is usual to state the phase of the current with respect to the voltage. For an inductive load the current lags the voltage and the power factor is said to be 'lagging'. For a predominantly capacitive load the current leads the voltage and the power factor is leading.

If the power is supplied from, say, an alternator rated at 400 V and 1000 A, then these are the highest voltage and current that the machine can tolerate without overheating. The phase difference between the voltage and current is entirely dependent upon the load. Thus if the power factor of the load is 1 then the 400 kV-A alternator can supply 400 kW of power to the load. Neglecting losses, the prime mover which drives the alternator must also be capable of supplying 400 kW. If, on the other hand, the power factor of the load is 0.5, then the power supplied will only be 200 kW. This means that, although the generator will be operating at its rated kV-A, the prime mover which drives the generator will be operating at only half its capacity.

An alternative way of looking at this phenomenon is to consider a load of, say, 100 kW, with a lagging power factor of 0.75. If the supply voltage is 50 V, then the required current (from equation (10.58)) is 2.67 A. If, however, the power factor of the load were to be increased to unity, then the required current would be reduced to 2 A. This means that the conducting cables, in supplying a reduced current, may have a correspondingly reduced cross-sectional area. In general, the size of an electrical system including transmission lines, switchgear and transformers is dependent upon the size of the current. It is economically viable, therefore, to ensure that the current is minimised. As a further incentive to industrial consumers, the electricity supply authorities normally operate a two-part tariff system. The two-part system consists of a fixed rate depending on the kV-A rating of the maximum demand and a running charge per kilowatts consumed per hour.

For these reasons it is advantageous to try and increase the power factor such that it is close to, but not quite, unity. A unity power factor is in fact avoided because it gives rise to a condition of resonance (see Section 10.2.30). In practice, capacitors, connected in parallel, are often used to improve the power factor of predominantly inductive loads such as electric motors. For large scale power systems a separate phase advance plant is used.

10.2.30 Frequency response of circuits

Since both inductive and capacitive reactance are frequency dependent, it is clear that the output from circuits involving capacitive and inductive elements will also depend on the frequency of the input signal.

The 'frequency response' of a circuit is usually presented as a plot of output/input ratio against the frequency as base. The ratio plotted could be one of voltages, currents or powers. Since the range of frequencies involved may be quite large, a logarithmic scale is normally employed. A logarithmic scale is also usually adopted for the vertical axis and the output/input ratio quoted in decibels (dB), i.e.

$$\text{Voltage ratio (in dB)} = 20\log_{10}\left[\frac{V_{out}}{V_{in}}\right] \quad (10.61)$$

Considering the series RLC circuit shown in *Figure 10.21*, and taking the voltage across the resistor as an output:

$$V_{out} = IR$$

$$V_{in} = I[R + j(\omega L - 1/\omega C)]$$

Therefore

$$\frac{V_{out}}{V_{in}} = \frac{R}{R + j(\omega L - 1/\omega C)}$$

Using the complex conjugate and calculating the modulus of the voltage ratio gives:

$$\left|\frac{V_{out}}{V_{in}}\right| = \frac{R}{[R^2 + (\omega L - 1/\omega C)^2]^{1/2}} \qquad (10.62)$$

The phase angle is given by

$$\phi = -\tan^{-1}\left[\frac{(\omega L - 1/\omega C)}{R}\right] \qquad (10.63)$$

It can be seen from equation (10.62) that the voltage ratio will have a maximum value of 1 when the frequency is such that $(\omega L - 1/\omega C) = 0$. Equating this expression gives:

$$\omega = \frac{1}{\sqrt{LC}} \qquad (10.64)$$

Equation (10.64) defines the so-called 'resonance' condition at which the inductive and capacitive reactances are equal and self-cancelling. The resonant frequency is usually denoted by ω_0 and it is the frequency at which the power transferred through the circuit is maximum. At any other frequency, above or below ω_0, the power transferred is reduced.

The impedance of the circuit is given by:

$$\bar{Z} = R + j(X_L - X_C) \qquad (10.65)$$

At the resonant frequency the total reactance is zero and the circuit behaves as if only the resistive element were present.

The general variation in the voltage ratio, or amplitude ratio, and phase angle with frequency is illustrated in *Figure 10.26*. Also shown on *Figure 10.26* are the two frequencies ω_1 and ω_2 at which the amplitude ratio is −3 dB. The −3 dB amplitude ratio is chosen because it corresponds to a halving in the power transmitted.

The 'bandwidth' is the frequency range between ω_1 and ω_2. A quality parameter, used with respect to resonant circuits, is the so-called 'Q factor' which is defined as the resonant frequency/bandwidth ratio.

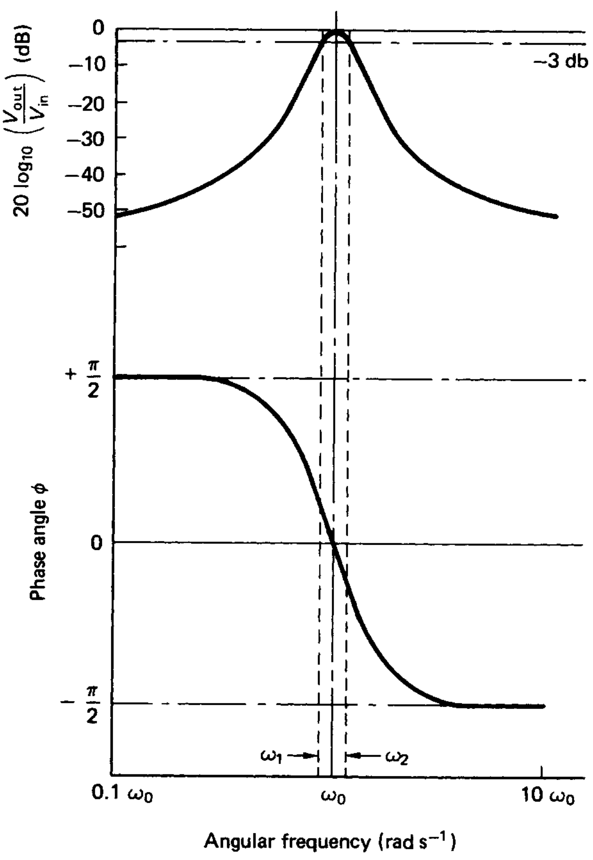


Figure 10.26 Voltage ratio and phase angle vs. frequency (series RLC)

10.2.31 Semiconductors

The common materials used for semiconductors are germanium and silicon. In recent times silicon has all but replaced germanium as a semiconductor material. The semiconductor materials have a crystalline structure such that each atom is surrounded by equally spaced neighbours. The basic structure can be visualised as a two-dimensional grid where the node points represent the central nucleus and the inner shell electrons, while the connecting lines of the grid represent the four valence electrons associated with each nucleus. This grid concept is adequate to describe an intrinsic, or 'pure', semiconductor.

At absolute zero temperature the crystalline structure is perfect and the electrons are all held in valence bonds. Since there are no current carriers available, the crystal behaves as a perfect insulator. As the temperature rises above absolute zero, an increasing number of valence bonds are broken, releasing pairs of free electrons and their associated 'holes'. In the absence of an applied field the free electrons move randomly in all directions. When an electric field is applied the electrons drift in a preferential direction to oppose the field and a net flow of current is established.

The covalent bond, with a missing electron, has a large affinity for electrons such that an electron from a neighbouring bond may easily be captured. This will leave the neighbouring atom depleted of electrons and the flow of electrons is generally associated with a counterflow of so-called 'holes'. The mobile hole, to all intents and purpose, is essentially a simple positive charge.

10.2.32 Doped semiconductors

Doped semiconductors are those in which an impurity has been introduced into a very pure intrinsic silicon. The nature of the impurity depends on the type of semiconductor required.

(1) *n type*—Impurities with five valence electrons can be added to produce a negative type of semiconductor. These impurities are referred to as 'donors' since the additional electron is very easily freed within the matrix. In the n-type semiconductor the free electrons are the dominant current carriers.

(2) *p type*—The p-type semiconductor is one in which the added impurities have only three valence electrons. Such impurities are called 'acceptors' and they produce a positive type of semiconductor within which hole conduction is the dominant current carrier.

10.2.33 p–n Junction diode

A p–n junction is formed by doping a crystal in such a way that the semiconductor changes from p-type to n-type over a very short length, typically 10^{-6} m. The transition zone from p-type to n-type is called the 'carrier depletion layer' and, due to the high concentrations of holes on one side and electrons to the other, a potential difference exists across this layer. The diffusion of holes from p to n and electrons from n to p is the majority carrier movement, called the 'diffusion current'. The drift of electrons from p to n and holes from n to p is the minority carrier movement referred to as the 'drift current'. When there is no externally applied potential difference, the diffusion current and the drift current are balanced in equilibrium. If an electric field is applied across the device, then two situations can exist, as illustrated in *Figure 10.27.*

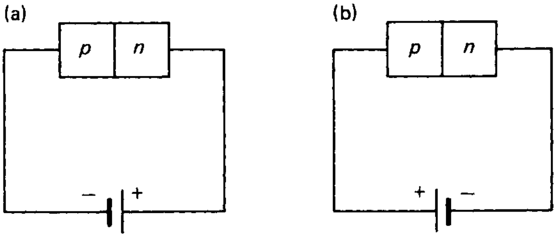


Figure 10.27 A p–n junction with applied potential difference. (a) Reverse bias; (b) forward bias

Figure 10.27(a) shows the reverse-bias mode in which the potential barrier is increased. The diffusion current is reduced while the drift current is barely altered. Overall, the current is negative and very small. When forward bias is applied (*Figure 10.27(b)*), the potential barrier is reduced and a large diffusion current flows. Overall, the current is positive and large. These general characteristics are the basis of a semiconductor diode which displays the typical current–voltage relationship depicted in *Figure 10.28.*

Figure 10.28 shows clearly that a very high impedance is presented by the diode to an applied voltage of reverse polarity. A low impedance is presented to a forward polarity voltage. In simple terms, the diode accommodates a forward flow of current, but greatly inhibits a reverse flow. The diode may be likened, therefore, to a switch which is activated 'on' for forward voltages and 'off' for reverse voltages. The reverse saturation current I_s is typically of the order of a few nanoamperes (nA) and can sensibly be regarded as zero.

The general characteristic also shows that the reverse voltage has a critical limiting value at which a 'breakdown' occurs.

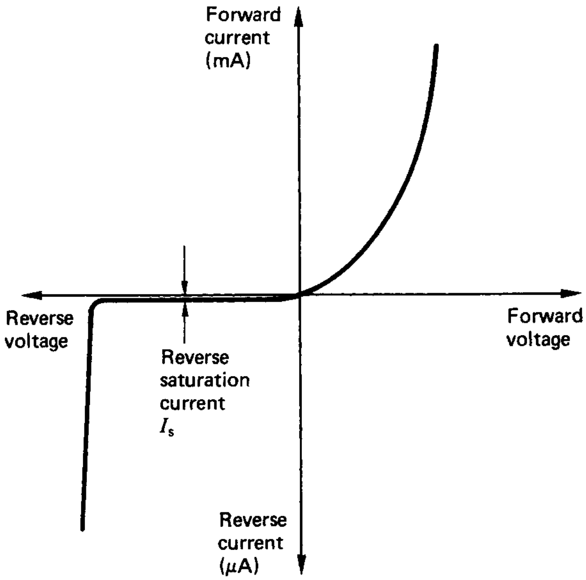


Figure 10.28 Current–voltage relationship for a p–n semiconductor diode

Depending upon the diode construction, the breakdown, or 'Zener' voltage may range from as low as 1 V to as much as several thousand volts. Up to the breakdown voltage, the reverse saturation current is independent of the reverse voltage.

Since the current–voltage relationship for a diode is a non-linear exponential function, the analysis of circuits involving diodes can become complicated. A simple awareness of the diode's practical function as a rectifier is perhaps more important than is a proficiency in analysing circuits involving diode elements.

10.2.34 A.c. rectification

Figure 10.29 shows an a.c. circuit with a diode in series with a load resistor. When the diode is forward biased, a current will flow in the direction indicated by the arrowhead. No current can flow when the diode is reverse biased, provided that the applied voltage does not exceed the breakdown value. The resultant current waveform through the resistor, for a sinusoidal voltage input, will therefore consist of positive only,

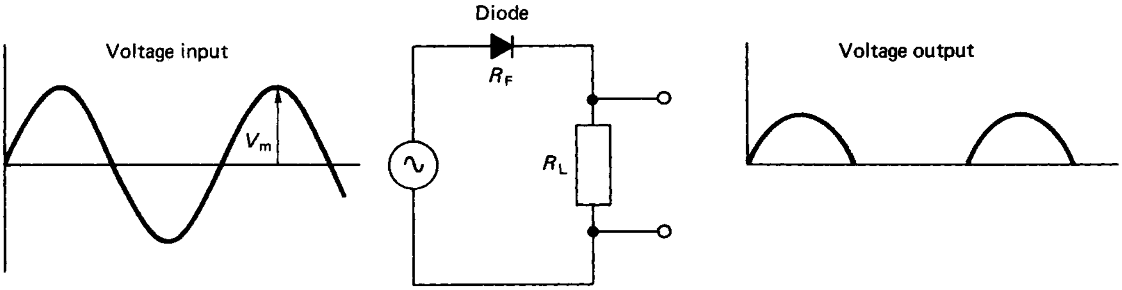


Figure 10.29 Half-wave rectification circuit

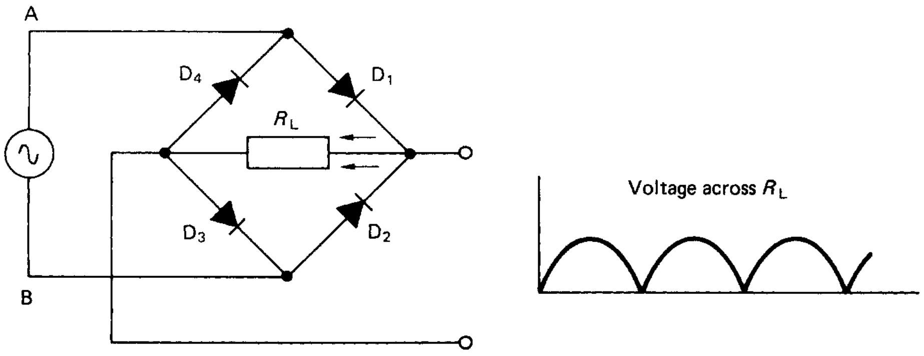


Figure 10.30 Full-wave rectification with a diode bridge

half sine waves. Since the output waveform is positive only, then it is by definition a d.c. voltage. It can be shown that the r.m.s. voltage across the resistor is:

$$V = \frac{V_m}{2} \frac{R_L}{R_L + R_F} \tag{10.66}$$

where R_L is the load resistance, R_F is the diode forward resistance, and V_m is the peak input voltage.

Determination of R_F is problematic, however, and models of varying complexity are used to simulate the diode in the circuit.

The single-diode circuit results in half-wave rectification. To obtain full-wave rectification, a diode bridge circuit can be used. The diode bridge is shown in *Figure 10.30*. When A is positive with respect to B then diodes D_1 and D_3 are conducting. When B is positive with respect to A then diodes D_2 and D_4 are conducting. The circuit arrangement ensures that the current, which consists of a continuous series of positive half sine waves, is always in the same direction through the load R_L.

With full-wave rectification there are twice as many half sine pulses through the load as there are with half-wave rectification. In addition, there are always two diodes effectively in series with the load. The resultant r.m.s. voltage across the load resistor for the full-wave diode bridge rectification circuit is:

$$V = \frac{V_m}{\sqrt{2}} \frac{R_L}{(R_L + 2R_F)} \tag{10.67}$$

The 'peak inverse voltage' (PIV) is defined as the maximum reverse biased voltage appearing across a diode. Thus when used as a rectifier, the diodes must have a sufficiently high reverse voltage rating in excess to the peak inverse voltage that the circuit can generate. For both the half- and the full-wave rectification circuits considered, the peak inverse voltage is equivalent to the maximum supply voltage V_m. Additional manufacturer's diode specifications would normally include the maximum power rating and the maximum allowable forward current.

10.2.35 The Zener diode

The diode breakdown effect is also used in a variety of circuits to provide a stabilised reference voltage. Special diodes which are designed to operate continuously in the reverse bias mode are called 'Zener diodes'. These diodes are manufactured with a range of breakdown voltages of 3–20 V. *Figure 10.31* shows

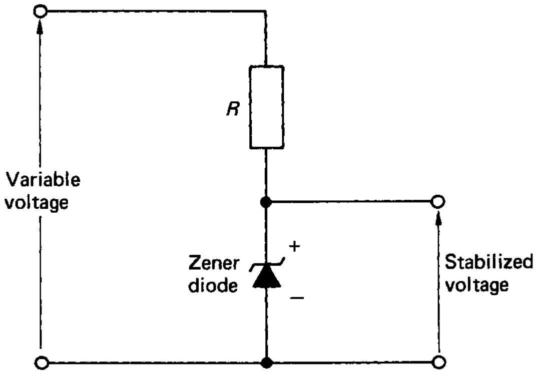


Figure 10.31 Zener diode as a reference voltage source

a Zener diode being used in a circuit to give a stable voltage which is essentially independent of the current flowing through the device. The series resistor in the circuit is included to limit the reverse current through the diode to a safe value.

10.3 Analogue and digital electronics theory

10.3.1 The bipolar, or junction, transistor

Originally developed in 1948, the transistor has revolutionised electronics and has influenced almost every aspect of life in the industrialised world. The term 'transistor', derived from 'transfer resistor', describes a device which can transfer a current from a low-resistance circuit to a high-resistance circuit with little change in current during the process.

The junction transistor consists of two p–n diodes formed together with one common section, making it a three-layer device (see *Figure 10.32*). Current flow in the transistor is both due to electron and hole conduction. The common central section is referred to as the 'base' and is typically of the order of 25 μm in length. Since the base can be made of either an n-type or a p-type semiconductor, then two basic configurations are possible. These are the n–p–n and the p–n–p types, as illustrated in *Figure 10.32*. The two other terminals are called

11 Microprocessors, Instrumentation and Control

C J Fraser

J S Milne

Contents

11.1 Basic control systems 11/3
- 11.1.1 Introduction 11/3
- 11.1.2 Mathematical models of systems—time domain analysis 11/5
- 11.1.3 Laplace notation for differential equations—frequency domain analysis 11/8
- 11.1.4 Stability criteria 11/9

11.2 Control strategies 11/13
- 11.2.1 ON/OFF control 11/13
- 11.2.2 Three-term, or PID, control 11/14
- 11.2.3 Empirical rules for PID controller settings 11/15
- 11.2.4 Three-term controller with a first-order system 11/16
- 11.2.5 Disturbance sensitivity 11/17
- 11.2.6 Direct digital control 11/17
- 11.2.7 Adaptive and self-tuning control 11/18
- 11.2.8 Sampled-data systems 11/20
- 11.2.9 Hierarchical control systems 11/21

11.3 Instrumentation and measurement 11/22
- 11.3.1 Analogue transducers 11/22
- 11.3.2 Digital transducers 11/23
- 11.3.3 Measurement of angular position 11/23
- 11.3.4 Measurement of linear displacement 11/25
- 11.3.5 Measurement of rotational speed 11/25
- 11.3.6 Transducers based on strain measurement 11/26
- 11.3.7 General force, pressure and acceleration transducers 11/28
- 11.3.8 Proximity sensing 11/29
- 11.3.9 Measurement of flow rate 11/29
- 11.3.10 Temperature measurement 11/31
- 11.3.11 Level sensing 11/31
- 11.3.12 Bar-code readers 11/32
- 11.3.13 Signal conditioning 11/33
- 11.3.14 Analogue and digital filtering 11/34

11.4 Microprocessor technology 11/36
- 11.4.1 Summary of number systems 11/36
- 11.4.2 ASCII code 11/36
- 11.4.3 Gray code 11/36
- 11.4.4 System architecture 11/37
- 11.4.5 Bus structure 11/37
- 11.4.6 Memory devices 11/38
- 11.4.7 Input/output structure 11/38
- 11.4.8 Memory map 11/39
- 11.4.9 Communication standards 11/40

11.5 Interfacing of computers to systems 11/41
- 11.5.1 Digital interfacing 11/41
- 11.5.2 Controller output interface hardware 11/42
- 11.5.3 Analogue interfacing 11/45
- 11.5.4 Multiplexing 11/48
- 11.5.5 Machine–tool interfaces 11/50
- 11.5.6 Robot control interfaces 11/50
- 11.5.7 Mobility and automated guided vehicles 11/52

11.6 Microprocessor-based control 11/52
- 11.6.1 Direct digital control 11/53
- 11.6.2 Hardware requirements 11/53
- 11.6.3 Software considerations 11/54
- 11.6.4 Sampling frequency in digital control loops 11/55
- 11.6.5 PID digital control algorithm 11/56
- 11.6.6 Speed control 11/57
- 11.6.7 The PC as a controller 11/58

11.7 Programmable logic controllers 11/59
- 11.7.1 The PLC in automation systems 11/59
- 11.7.2 The PLC vs. the microcomputer 11/60
- 11.7.3 Ladder logic programming 11/60
- 11.7.4 Controlling pneumatic and hydraulic systems 11/62
- 11.7.5 Safety 11/63
- 11.7.6 Networking of PLCs 11/63

11.8 Robot applications 11/64
11.8.1 Robot geometry 11/65
11.8.2 Robot programming 11/67
11.8.3 Path control 11/68
11.8.4 Computational hardware 11/68

11.1 Basic control systems

11.1.1 Introduction

Control engineering is based on the linear systems analysis associated with the development of feedback theory. A control system is constituted as an interconnection between the components which make up the system. These individual components may be electrical, mechanical, hydraulic, pneumatic, thermal or chemical in nature, and the well designed control system will provide the 'best' response of the complete system to external, time-dependent disturbances operating on the system. In the widest sense, the fundamentals of control engineering are also applicable to the dynamics of commercial enterprise, social and political systems and other non-rigorously defined concepts. In the engineering context, however, control principles are more generally applied to much more tangible and recognisable systems and subsystems.

Invariably, the system to be controlled can be represented as a block diagram, as in *Figure 11.1*. The system is a group of physical components combined to perform a specific function. The variable controlled may be temperature, pressure, flow rate, liquid level, speed, voltage, position, or perhaps some combination of these. Analogue (continuous) or digital (discrete) techniques may individually, or simultaneously, be employed to implement the desired control action. More recently, the advances made in microelectronics have resulted in a growing emphasis on digital techniques, and the majority of modern control systems are now microprocessor based.

11.1.1.1. Classification of control systems

Engineering control systems are classified according to their application and these include the following.

Servomechanisms Servomechanisms are control systems in which the controlled variable, or output, is a position or a speed. D.c. motors, stepper motor position control systems and some linear actuators are the most commonly encountered examples of servomechanisms. These are especially prevalent in robotic arms and manipulators.

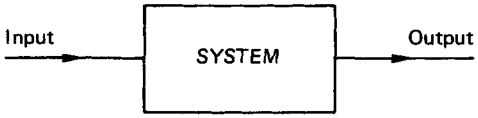


Figure 11.1 System to be controlled

Sequential control A system operating with sequential control is one where a set of prescribed operations is performed in sequence. The control may be implemented as 'event based', where the next action cannot be performed until the previous action is completed. An alternative mode of sequential control is termed 'time based', where the series of operations is sequenced with respect to time. Event-based sequential control is intrinsically a more reliable 'fail-safe' mode than the time-based type. Consider, for example, an industrial process in which a tank is to be filled with a liquid and the liquid subsequently heated. The two control systems are depicted in *Figure 11.2*.

The time-based sequential control system (*Figure 11.2(a)*) is the simplest. The pump is switched on for an interval which would discharge enough liquid into the tank to fill it to approximately the correct level. Following this, the pump is switched off and the heater is switched on. Heating is similarly allowed to continue for a preset time, after which the liquid temperature would approximately have reached the desired value. Note that the control function is inexact and there are no fail-safe features. If the drive shaft between the motor and the pump becomes disengaged, or broken, then the heater will still come on at the prescribed time, irrespective of whether there is liquid in the tank or not. The event-based sequential control system has fail-safe features built-in and is much more exact. In operation the pump is switched on until the liquid-level sensor indicates that the tank is filled. Then, and only then, is the pump switched off and the heater switched on. The temperature of the liquid is also monitored with a sensor and heating is applied until such time that the temperature reaches the desired value. Obviously, with two additional sensors, the event-based system is the more expensive. The advantages it offers over the time-based system, however, far outweigh its disadvantages, and event-based sequentially controlled systems are by far the most common. Time-based systems do exist nonetheless, and they are found in applications where the results of malfunction would be far less potentially catastrophic than those occurring in the example described. The essential difference between the two systems is that event-based sequential control incorporates a check that any operation has been completed before the next is allowed to proceed. The modern automatic washing machine and automatic dishwasher are examples of sequentially controlled systems.

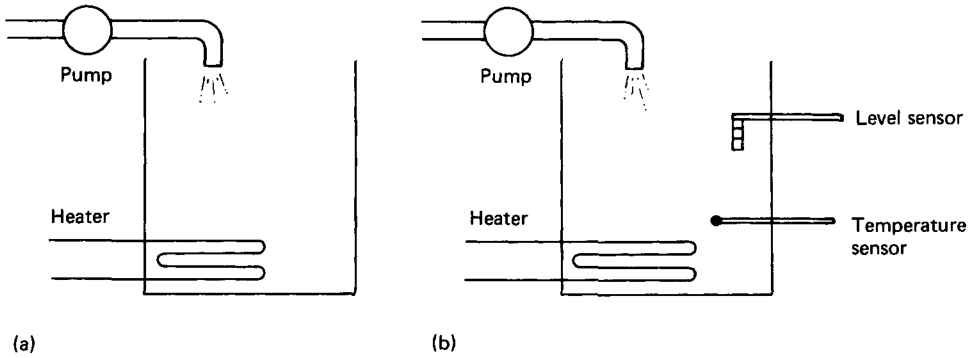


Figure 11.2 Simple sequential control systems: (a) time based; (b) event based

Numerical control In a system using numerical control, the numerical information is stored in the form of digital codes on a control medium which may be a paper tape, a magnetic-sensitive tape, or a magnetic-sensitive disc. This information is used to operate the system in order to control such variables as position, direction, velocity and speed. There is a wide variety of manufacturing operations involving machine tools which utilise this versatile method of control.

Process control In this type of control the variables associated with any process are monitored and subsequent control actions are implemented to maintain the variables within the predetermined process constraints. The word 'process' is all-encompassing and might include, for example, electrical power generation. The generation of 'electricity' can be considered as a manufacturing process where the 'product' is kilowatt-hours. In the control of power generation, the variables which are measured include temperature, pressure, liquid level, speed, flow rate, voltage, current and a range of various gas concentrations. This is further complicated by the need to satisfy the power demand, and it is apparent that the control of such a system is necessarily complex. Similarly complex examples exist in the oil and paper making industries, in the automotive assembly industries and in any entity which aspires to the designation of a 'flexible manufacturing system'.

11.1.1.2 Open- and closed-loop control

The basic open-loop system is shown in *Figure 11.1* and extended in *Figure 11.3* to give a more complete picture. The input element supplies information regarding the desired value X of the controlled variable. This information is then acted on by the controller to alter the output Y through the load. External disturbances are fed in as shown and will cause the output to vary from the desired value. The open-loop system may be likened to the driving of a vehicle where the driver constitutes the input element. Essentially, two variables are controlled by the driver and they are the speed and the direction of motion of the vehicle. The controller, in the case of speed, is the engine throttle valve and in the case of direction, is the steering system.

In order for the system to become a closed loop, two further elements must be added:

(1) a monitoring element to measure the output Y; and
(2) a comparing element to measure the difference between the actual output and the desired value X.

The monitoring and comparing elements are connected through the 'feedback' link, as shown in *Figure 11.4*.

It is a moot point and can be argued that the driver in the above example also performs the functions of monitoring and comparing. The vehicle driver, therefore, if considered to be part of the complete system, constitutes a closed-loop feedback control. For the purpose of definition, however, any system which incorporates some form of feedback is termed a 'closed-loop system'. With no feedback mechanism, the system is categorised as 'open loop'. For the most practical engineering purposes, control systems are of the closed-loop variety to take advantage of the benefits of feedback, which may be either 'positive' or 'negative'. A positive feedback signal aids the input signal. It is possible, therefore, to have output with no input when using a positive-feedback signal and, since this is detrimental to the control function, positive-feedback systems are very rare.

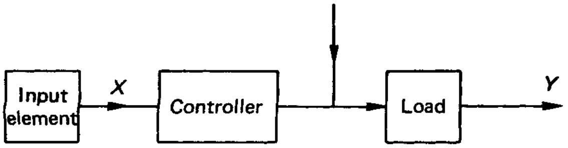


Figure 11.3 Open-loop control system

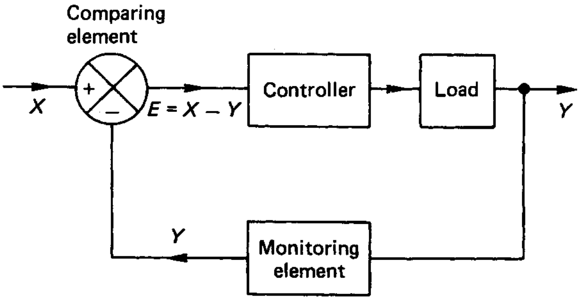


Figure 11.4 Closed-loop feedback control system

11.1.1.3 Linear and non-linear control systems

For a control system to be linear it must satisfy both the amplitude proportionality criteria and the principle of superposition. If a system output at a given time is $Y(t)$ for a given input $X(t)$, then an input of $kX(t)$ must produce an output of $kY(t)$ if amplitude proportionality is satisfied. In a similar manner, if an input of $X_1(t)$ produces an output of $Y_1(t)$, while an input of $X_2(t)$ produces an output of $Y_2(t)$ then, if an input of $[X_1(t) + X_2(t)]$ produces an output of $[Y_1(t) + Y_2(t)]$ the superposition principle is satisfied. Non-linear systems do not necessarily satisfy both these criteria and these systems are generally 'compensated' such that their behaviour approaches that of an equivalent linear system.

11.1.1.4 Characteristics of control systems

The characteristics of a control system are related to the output behaviour of the system in response to any given input. The parameters used to define the characteristics of a control system are: stability, accuracy, speed of response and sensitivity.

The system is said to be 'stable' if the output attains a certain value in a finite interval after the input has undergone a change. When the output reaches a constant value the system is said to be in the 'steady state'. The system is unstable if the output increases with time. In any practical control system, stability is absolutely essential. Systems involving a 'time delay', or a 'dead time' may tend to be unstable, and extra care must be taken in the design of such systems to ensure stability. The stability of control systems can be analysed using various analytical and graphical techniques. These include the Routh–Hurwitz criteria and the Bode, Nichols and Nyquist graphical methods.

The accuracy of a system is a measure of the deviation of the actual controlled value in relation to its desired value. Accuracy and stability are interactive and one can in fact be counter-productive to the other. The accuracy of a system might be improved but, in refining the limits of the desired output, the stability of the system might be adversely affected. The converse also applies.

The speed of response is a measure of how quickly the output attains a steady-state value after the input has been altered.

Sensitivity is an important factor and is a measure of how the system output responds to external environmental conditions. Ideally, the output should be a function only of the input and should not be influenced by undesirable extraneous signals.

11.1.1.5 Dynamic performance of systems

The dynamic performance of a control system is assessed by mathematically modelling, or experimentally measuring, the output of the system in response to a particular set of test input conditions.

(1) *Step input*—This is perhaps the most important test input, since a system which is stable to a step input will also be stable under any of the other forms of input. The step input (*Figure 11.5*) is applied to gauge the transient response of the system and gives a measure of how the system can cope with a sudden change in the input.
(2) *Ramp input*—A ramp input (*Figure 11.6*) is used to indicate the steady-state error in a system attempting to follow a linearly increasing input.

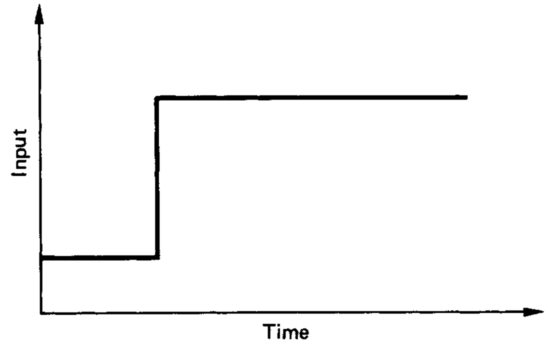


Figure 11.5 Step input

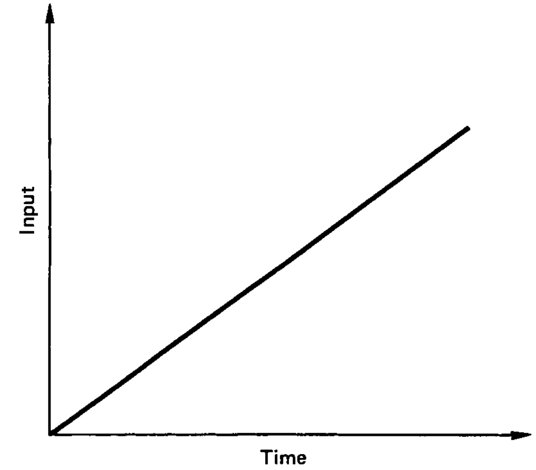


Figure 11.6 Ramp input

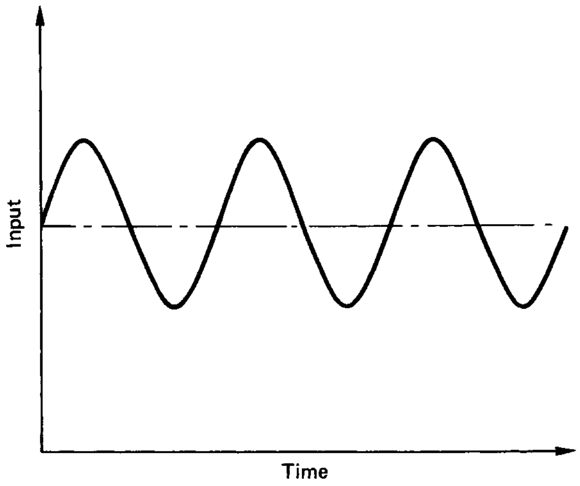


Figure 11.7 Sinusoidal input

(3) *Sinusoidal input*—The sinusoidal input (*Figure 11.7*) over a varying range of input frequencies, is the standard test input used to determine the frequency response characteristics of the system.

Although the three standard test inputs may not be strict representations of the actual inputs to which the system will be subject, they do cover a comprehensive range. A system which performs satisfactorily under these inputs will, in general, perform well under a more natural range of inputs. The system response to a parabolically varying test input can also be analysed, or measured, but this is a less commonly used test signal compared with the above three.

11.1.1.6 Time domain and frequency domain

The time domain model of a system results in an output $Y(t)$ with respect to time for an input $X(t)$. The time domain system model is expressed as a differential equation, the solution of which is displayed as a graph of output against time.

In contrast, a frequency domain model describes the system in terms of the effect that the system has on the amplitude and phase of sinusoidal inputs. Typically, the system performance is displayed in plots of amplitude ratio $[Y(t)/X(t)]$ or $20 \log_{10}[Y(t)/X(t)]$, and phase angle, against input signal frequency.

Neither system model has an overriding advantage over the other, and both are used to good effect in describing system performance and behaviour.

11.1.2 Mathematical models of systems—time domain analysis

Differential equations are used to model the relationship between the input and output of a system. The most widely used models in control engineering are based on first- or second-order linear differential equations.

11.1.2.1 First-order systems

Some simple control systems, including those for the control of temperature, level and speed, can be modelled as a first-order linear differential equation:

$$\tau \frac{dY}{dt} + Y = kX \qquad (11.1)$$

where X and Y are the input and output, respectively, τ denotes the system time constant, and k is the system gain. When the input X is a step of amplitude A then the solution to equation (11.1) gives the result shown in *Figure 11.8*. The solution curve shown in *Figure 11.8* has the analytical form:

$$Y(t) = kA[1 - e^{-t/\tau}] \qquad (11.2)$$

Equation (11.2), which is the time-domain solution, is an exponential function which approaches the value kA as t approaches infinity. Theoretically, the output never reaches kA and the response is termed an 'exponential lag'. The time constant τ represents the time which the output would take to reach the value kA if the initial rate of response were maintained. This is indicated in *Figure 11.8* by the broken line which is tangential to the solution curve at time $t = 0$. For practical purposes, the final steady-state output is taken to have been reached in a time of about 5τ.

If the input is a ramp function, then the response of a first-order system is as shown in *Figure 11.9*. The ramp input is simulated by making the right-hand side of equation (11.1) a linear function of time, i.e. kAt. With this input, the time domain solution becomes:

$$Y(t) = kA[t - \tau(1 - e^{-t/\tau})] \qquad (11.3)$$

The solution equation shows that, as t becomes large, the output tends to $kA(t - \tau)$. Therefore the output response is asymptotic to a steady-state lag $kA\tau$.

The response of a first-order system to a sinusoidal input can be obtained by setting the right-hand side of equation (11.1) equal to $kA \sin(\omega t)$, where ω is a constant circular frequency (in rad s^{-1}). The time domain solution yields:

$$Y(t) = \frac{kA}{\sqrt{(1 + \tau^2\omega^2)}} [\sin \alpha e^{-t/\tau} + \sin(\omega t - \alpha)] \qquad (11.4)$$

where $\alpha = \tan^{-1}(\tau\omega)$.

The response is shown in *Figure 11.10*. The output response exhibits a decaying transient amplitude in combination with a steady-state sinusoidal behaviour of amplitude, $kA[\sqrt{(1 + \tau^2\omega^2)}]$ and lagging the input by the angle α.

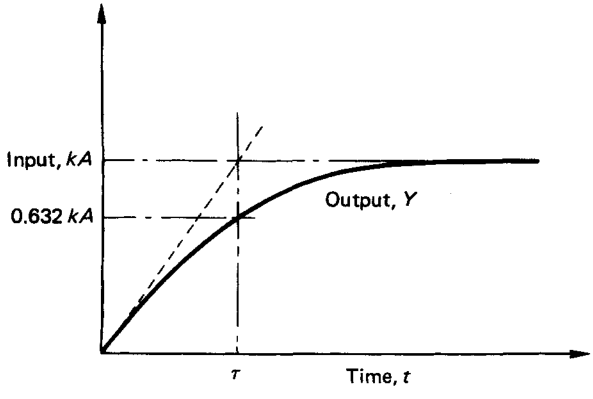


Figure 11.8 Response of a first-order system to a step input

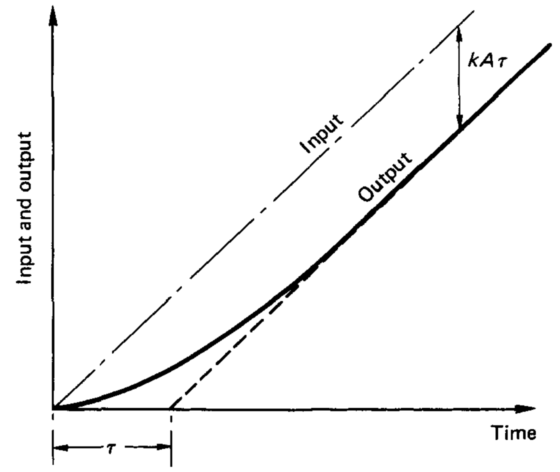


Figure 11.9 First-order system response to a ramp input

11.1.2.2 Second-order systems

While some control systems may be adequately modelled as a first-order linear differential equation, many more practical systems, including position control, are more usually represented by a differential equation of second order. The second-order differential equation has the general form:

$$\frac{d^2Y}{dt^2} + 2\zeta\omega_n \frac{dY}{dt} + \omega_n^2 Y = kX \qquad (11.5)$$

where ζ is termed the 'damping ratio' and is defined as the ratio of the actual damping in the system to the damping which would produce critical damping. ω_n is the undamped natural frequency of the system, and k is the system gain.

The time-domain solution depends on the magnitude of ζ and three solutions for a step input are possible.

(1) *Light damping, $\zeta < 1$*:

$$Y(t) = \exp(-\zeta\omega_n t)[A \cos(\omega_n\sqrt{1 - \zeta^2}t) + B \sin(\omega_n\sqrt{1 - \zeta^2}t)] \qquad (11.6)$$

(2) *Critical damping, $\zeta = 1$*:

$$Y(t) = \exp(-\zeta\omega_n t)[At + B] \qquad (11.7)$$

(3) *Heavy damping, $\zeta > 1$*:

$$Y(t) = A\exp[-\zeta\omega_n t - (\omega_n\sqrt{\zeta^2 - 1})t] + B \exp[-\zeta\omega_n t + (\omega_n\sqrt{\zeta^2 - 1})t] \qquad (11.8)$$

The three possible solutions, for a step input, are shown in *Figure 11.11*, where the output $Y(t)$ is plotted as a percentage of the step input X against the parameter $(\omega_n t)$.

For $\zeta = 1$, the system is critically damped and the steady-state value is attained in the shortest possible time without any oscillatory response. With $\zeta > 1$, the system is overdamped and the response curve is again exponential in form. Overdamped systems may have an undesirably sluggish response. Indeed, since the effect of $\zeta > 1$ simply delays the response to the steady-state value, then there is no real advantage to be gained in using high ζ values. For cases where $\zeta < 1$, the system is said to be 'underdamped' and the response curve is oscillatory with an exponential decay. A number of perfor-

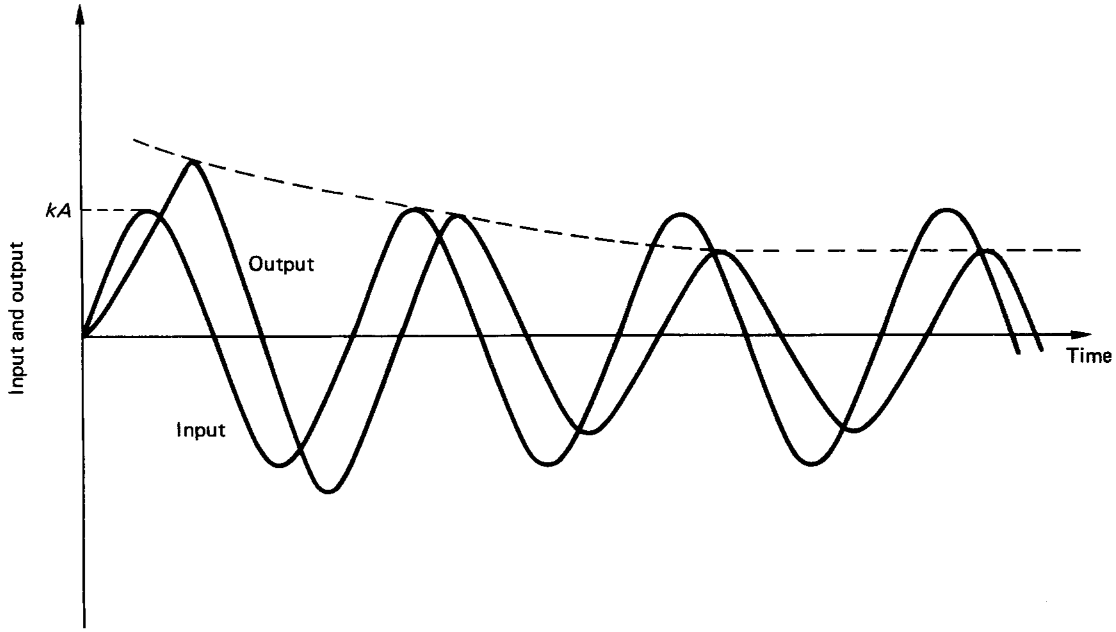


Figure 11.10 First-order system response to a sinusoidal input

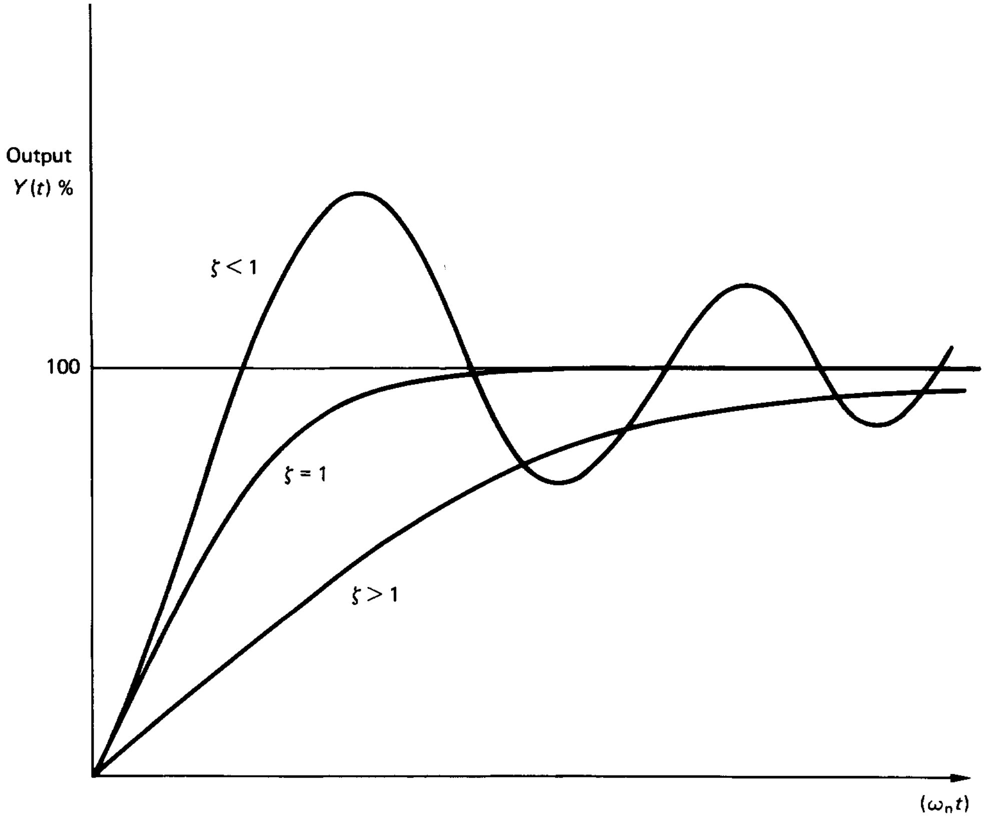


Figure 11.11 Response of a second-order system to a step input

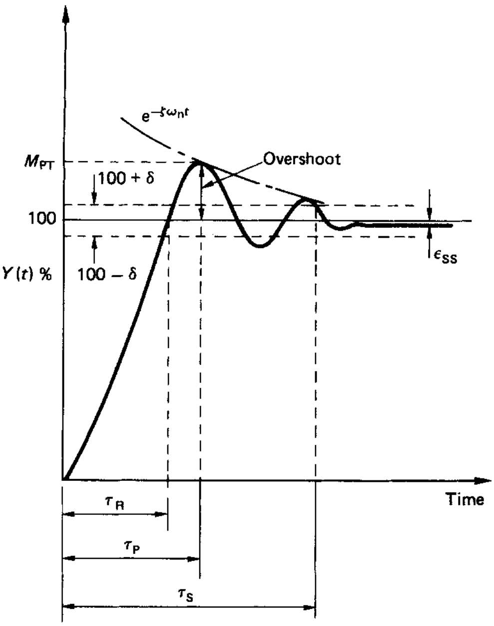


Figure 11.12 Response curve for an underdamped system to a step input

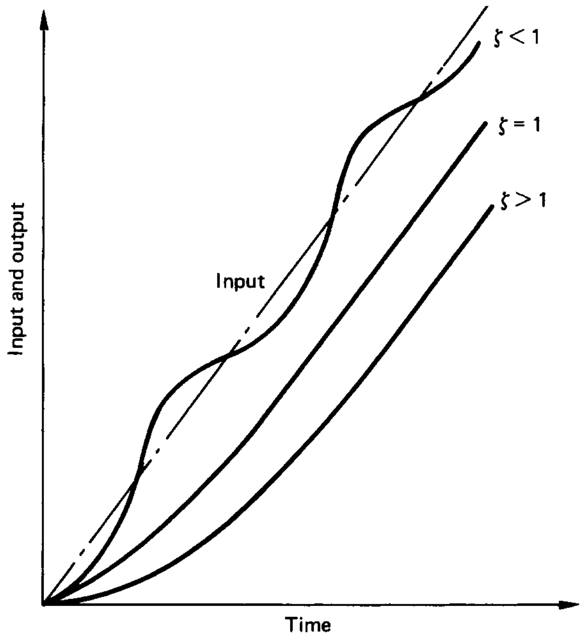


Figure 11.13 Response of a second-order system to a ramp input

mance measures are used to describe the response of an underdamped system to a step input. These are illustrated in *Figure 11.12*.

The speed of the response is reflected by the rise time τ_R and the peak time τ_p. For underdamped systems, the rise time is the time taken for the output to reach 100% of the step input. The peak time is the time taken to the first maximum in the output response. For critically damped and overdamped systems, the time taken for the output to change between 10% and 90% of the input is used as a measure of the speed of the response.

The degree in which the actual output response matches the input is measured by the percentage overshoot (PO) and the settling time. The percentage overshoot is defined as:

$$PO = \frac{M_{PT} - 100}{100} \qquad (11.9)$$

where M_{PT} is the peak value of the output.

It may further be shown that the percentage overshoot is given analytically as:

$$PO = 100\,e^{-\zeta\pi/\sqrt{1-\zeta^2}} \qquad (11.10)$$

Another useful relationship is derived from the ratio of successive peaks, i.e.

$$\ln\left(\frac{A_n}{A_{n+1}}\right) = \frac{2\zeta\pi}{\sqrt{1-\zeta^2}} \qquad (11.11)$$

where n is an integer denoting the peak number, i.e. first, second, etc. Equation (11.11) is referred to as the 'logarithmic decrement'.

The settling time τ_s is the time taken for the oscillatory response to decay below a percentage of the input amplitude δ, which is often taken as ±2%.

Finally, we have the steady-state error ε_{ss} which is self-explanatory.

The response of the second-order system to a ramp input is shown in *Figure 11.13*. The form of the response curves again depends on the value of the damping ratio, but in each case the output asymptotes to a steady-state lag. The lag is not the same in each case, however, since it is also dependent on the damping ratio.

The response of a second-order system to a sinusoidal input may also be considered. In general, the output response will lag behind the input with a transient decaying amplitude depending on the nature of the damping ratio. It is more informative, however, to study the response of second-order systems to sinusoidal inputs using frequency domain methods.

11.1.3 Laplace notation for differential equations—frequency domain analysis

For analyses in the frequency domain it is customary to write the differential equation in terms of the Laplace operator s. This gives rise to the system 'transfer function' which is formed by replacing the input and output (X and Y, respectively) with their corresponding Laplace transforms ($X(s)$ and $Y(s)$). The method applies only to linear differential equations. In practice, many systems would contain some degree of non-linearity, and various assumptions would have to be made to simplify and approximately linearise the governing equation.

The advantage of using the Laplace transform method is that it allows the differential equation to be expressed as an equivalent algebraic relation in s. Differentiation is represented by multiplication by the Laplace variable s. Thus dY/dt becomes $sY(s)$ and d^2Y/dt^2 becomes $s^2Y(s)$.

11.1.3.1 First-order systems

The governing equation is rewritten with the appropriate Laplace transforms replacing the differential operators. Thus equation (11.1) becomes:

$$\tau sY(s) + Y(s) = kX(s) \quad (11.12)$$

Hence,

$$Y(s)[1 + \tau s] = kX(s) \quad (11.13)$$

The system transfer function is defined as the ratio of the output to the input:

$$\frac{Y(s)}{X(s)} = \frac{k}{1 + \tau s} \quad (11.14)$$

Equation (11.14) enables the convenient facility of incorporating the transfer function within the usual block-structure representation of a control system. Thus a first-order, open-loop control system can be systematically depicted as shown in *Figure 11.14*.

For analyses in the frequency domain, we are predominantly concerned with the system response to sinuosidal inputs. Differentiation, or integration, of a sinusoidal function does not alter the shape or frequency. There is simply a change in amplitude and phase, e.g.

$$\text{Input} = A \sin(\omega t) \quad (11.15)$$

$$\frac{d}{dt}[A \sin(\omega t)] = \omega A \cos(\omega t)$$

$$= \omega A \sin\left(\omega t + \frac{\pi}{2}\right) \quad (11.16)$$

Comparison of equations (11.15) and (11.16) shows that differentiation has changed the amplitude from A to ωA and that there is a phase shift of 90° associated with the process.

Equation (11.16) thus describes the steady-state output from a first-order, open-loop control system. The transient part of the output, which the time-domain solution illustrates (*Figure 11.10*), is not apparent in the frequency-domain solution.

The Laplace operator s may be replaced with $j\omega$, where j is the complex operator $\sqrt{-1}$. Equation (11.14) then becomes

$$\frac{Y}{X} = \frac{k}{1 + j\omega\tau}$$

Using the complex conjugate, it may be shown that the modulus of the amplitude ratio is

$$\left|\frac{Y}{X}\right| = \frac{k}{\sqrt{1 + \omega^2\tau^2}} \quad (11.17)$$

Equation (11.17) shows how the output amplitude will be influenced by the input sinusoidal frequency. Note that this agrees with the time domain solution (equation (11.4)) for large values of t, after which the steady-state is achieved.

The technique described above is general and may be used to determine the amplitude ratio for any second, or higher, order system.

N.B. Common practice, especially in graphical representations, is to express the amplitude ratio in decibels, i.e.

$$\text{dB} = 20 \log_{10}(Y/X) \quad (11.18)$$

11.1.3.2 Second-order systems

Using the Laplace transfer operator, the governing equation (equation (11.5)) may be rewritten as

$$s^2Y(s) + 2\zeta\omega_n sY(s) + \omega_n^2 Y(s) = kX(s) \quad (11.19)$$

Therefore,

$$Y(s)[s^2 + 2\zeta\omega_n s + \omega_n^2] = kX(s) \quad (11.20)$$

The system transfer function is

$$\frac{Y(s)}{X(s)} = \frac{k}{s^2 + 2\zeta\omega_n s + \omega_n^2} \quad (11.21)$$

Thus a second-order, open-loop control system can be represented schematically as shown in *Figure 11.15*.

For a sinusoidal input of the form $X = X_0 \sin(\omega t)$, the frequency-domain analysis gives the following steady-state solutions for the amplitude ratio and the phase lag:

$$\left|\frac{Y}{X}\right| = \frac{1}{\sqrt{(1 - r^2)^2 + (2\zeta r)^2}} \quad (11.22)$$

and

$$\phi = \tan^{-1}[(2\zeta r)/(1 - r^2)] \quad (11.23)$$

where $r = (\omega/\omega_n)$.

The above frequency-response characteristics are shown in *Figure 11.16*.

When the input signal frequency is equal to the system's natural frequency, the amplitude ratio has the value of $(1/2\zeta)$ and the phase lag is −90°. Note that, if the damping ratio is zero, then the amplitude ratio theoretically approaches infinity under this resonance condition. In practice, if the damping ratio is moderately low, then very large output amplitudes can be expected if the input frequency is in the vicinity of the natural frequency of the system.

Thus far we have considered the open-loop system response for first- and second-order systems. Such systems are unconditionally stable. The addition of a feedback loop, however, increases the order of the system, and there is always the possibility that the second-order system with feedback may be unstable. Furthermore, if any system, either first or second order, incorporates a 'time delay' (also known as a 'deadtime' or a 'transportation lag'), then unstable operation is more likely to occur.

11.1.4 Stability criteria

Time delays are very difficult to handle mathematically when they occur in differential equations, and the inclusion of

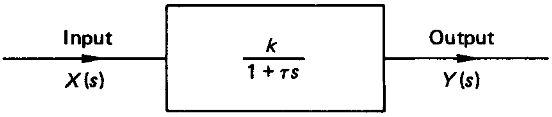


Figure 11.14 First-order, open-loop control system

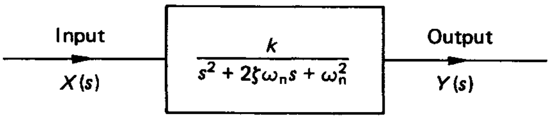


Figure 11.15 Second-order, open-loop control system

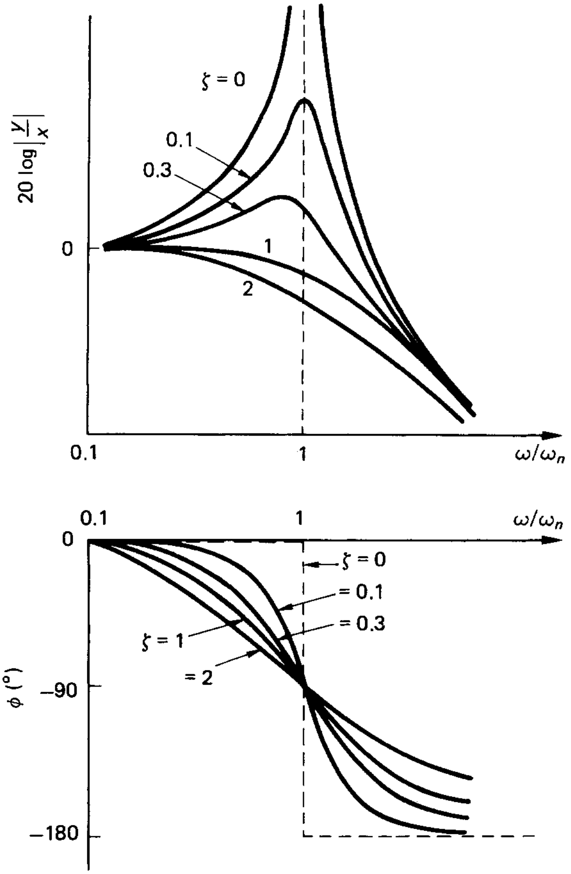


Figure 11.16 Second-order, open-loop control system frequency response

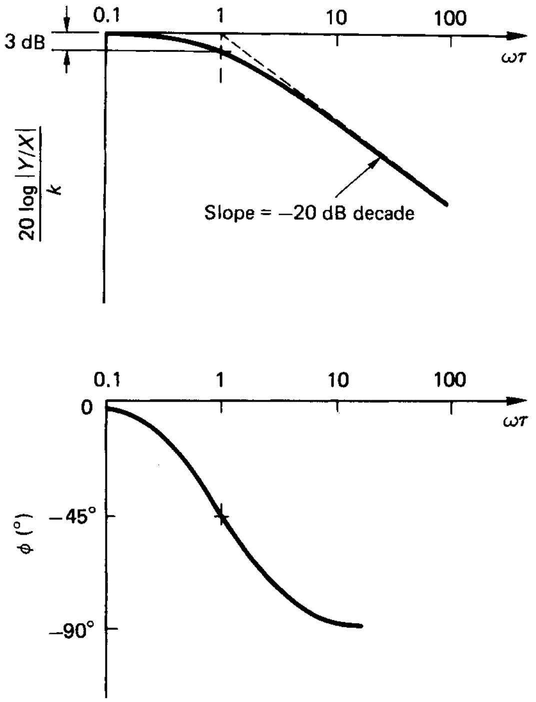


Figure 11.17 Bode plot for an open-loop, first-order system

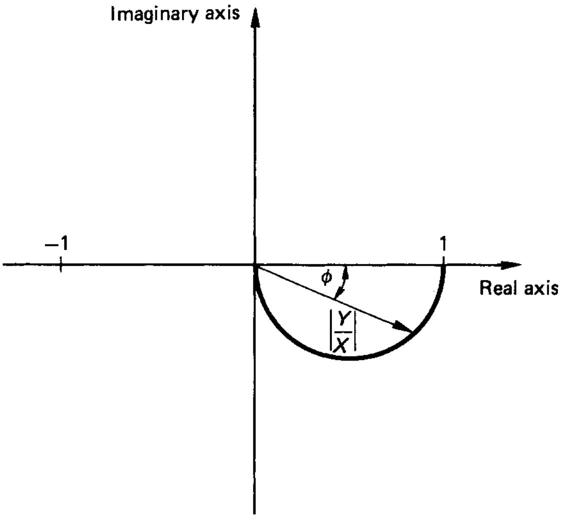


Figure 11.18 Nyquist plot for an open-loop, first-order system

multiple feedback loops can greatly increase the order of the governing equation. For these two reasons solutions in the time domain become extremely difficult, and frequency-domain methods are almost exclusively used to assess the behaviour of more complex control systems. The main consideration in frequency-domain analyses is the stability of the system and how the system can be adjusted if it happens to be unstable. Various graphical methods are used and these include the Bode and Nyquist plots.

The Bode plot is a graph of amplitude ratio and phase angle variation with input signal frequency. The resulting normalised plot for an open-loop, first-order system is shown in *Figure 11.17*. Note that, when the input frequency is equal to the inverse of the system time constant, the output amplitude has been decreased, or attenuated, by 3 dB. The phase lag at this point is −45°. This is characteristic of first-order systems.

The Nyquist plot represents the same information in an alternative form. The plot is in polar coordinates and combines the amplitude ratio and phase lag in a single diagram. *Figure 11.18* shows the Nyquist plot for the open-loop, first-order system.

11.1.4.1 Bode and Nyquist stability criteria

The Bode[1] criterion for stability is: *the system is stable if the amplitude ratio is less than 0 dB when the phase angle is −180°*

This is illustrated graphically in *Figure 11.19*, which represents a stable system because Bode's criterion is satisfied. The 'gain margin' (GM) and 'phase margin' (PM) are used as measures of how close the frequency-response curves are to the 0 dB and −180° points, and are indicative of the relative stability of the system.

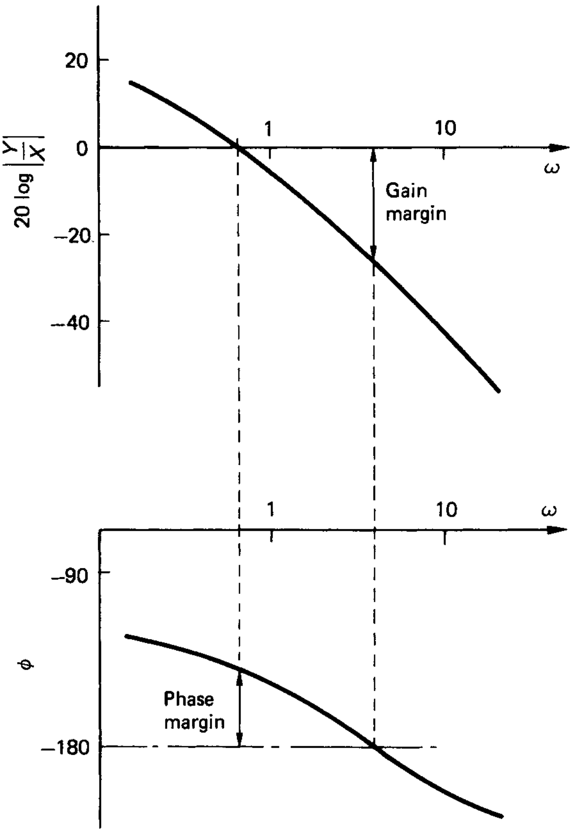


Figure 11.19 Bode's stability criterion

The Nyquist[2] criterion for stability is: *the system is stable if the amplitude ratio is greater than −1 at a phase angle of −180°*. In effect, this means that the locus of the plot of amplitude ratio and phase angle must not enclose the point −1 on the real axis. A stable response curve is shown plotted in *Figure 11.20*. Also indicated in *Figure 11.20* are the gain margin and phase margin in the context of the Nyquist plot.

11.1.4.2 System stability with feedback

In a closed-loop system, the transfer function becomes modified by the feedback loop. The first task, therefore, is to determine the overall transfer function for the complete system.

For simple open-loop systems the transfer functions are combined according to the following rules:

(1) for elements in series, the overall transfer function is given by the product of the individual transfer functions (see *Figure 11.21*); and
(2) for elements in parallel, the overall transfer function is given by the sum of the individual transfer functions (see *Figure 11.22*).

For a system with feedback, the overall transfer function can be evaluated using a consistent step-by-step procedure. Series and parallel control elements are combined in the manner shown above in order to reduce the system to a single block which then represents the overall transfer function.

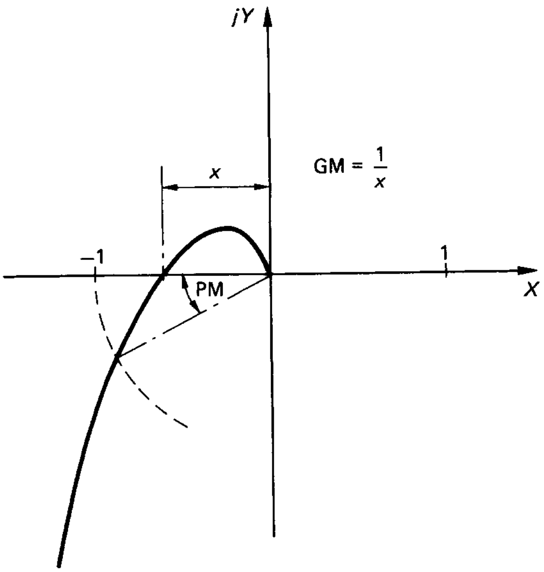


Figure 11.20 Nyquist's stability criterion

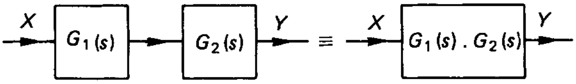


Figure 11.21 Transfer functions in series

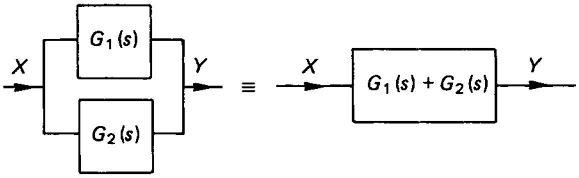


Figure 11.22 Transfer functions in parallel

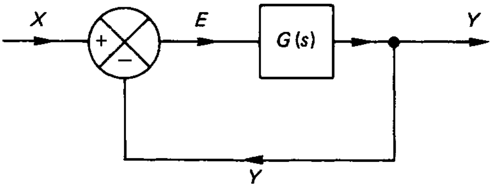


Figure 11.23 Control system with unity feedback

Consider the simple control system depicted in *Figure 11.23*. Since the feedback line does not include any transfer function, it is termed a 'unity feedback' system, i.e. the output is compared directly with the input to produce the error signal. The closed-loop transfer function is obtained as follows:

$Y = G(s)E = G(s)[X - Y]$

Thus

$$\frac{Y}{X} = \frac{G(s)}{1 + G(s)} \quad (11.24)$$

If the element whose open-loop transfer function $G(s)$ is a first-order subsystem, then $G(s)$ may be replaced by the expression given in equation (11.14). The closed-loop transfer function may then be written as

$$\frac{Y}{X} = \left(\frac{k}{1 + \tau s}\right) \bigg/ \left(1 + \frac{k}{1 + \tau s}\right)$$

$$= \frac{k}{1 + k + \tau s} \quad (11.25)$$

Dividing top and bottom by $(1 + k)$ results in

$$\frac{Y}{X} = \left(\frac{k}{1 + k}\right) \bigg/ \left(1 + \frac{\tau s}{1 + k}\right) \quad (11.26)$$

Defining the following terms

$$k_c = \frac{k}{1 + k} \quad (11.27)$$

and

$$\tau_c s = \frac{\tau s}{1 + k} \quad (11.28)$$

where k_c is the closed-loop gain and τ_c is the closed-loop system time constant. The final closed-loop transfer function may be expressed as

$$\frac{Y}{X} = \frac{k_c}{1 + \tau_c s} \quad (11.29)$$

Equations (11.27) and (11.28) show, respectively, that both the closed-loop system gain and the time constant are less than those associated with the open-loop system. This means that the closed-loop response is faster than the open-loop response. At the same time, however, the closed-loop gain is reduced.

Using the procedures outlined in the example, any other complex control system may be similarly analysed to determine the closed-loop transfer function. Thus, knowing the gain constants and other characteristics of the elements which make up the system, the frequency response may be obtained. The stability of the system may then be assessed and any corrective measures taken as necessary. In practice, it is often found that the gain of some of the system elements must be altered in order to ensure stable operation. Another commonly applied corrective measure is to add a phase advance circuit into the system. The procedure might also be operated in reverse—by starting with a desired response, a suitable control system can be configured and adjusted to meet the response.

The practising control engineer will use many techniques to assess system stability. These might include the numerical Routh–Hurwitz criterion, which determines only whether a system is stable or not. Alternative graphical methods include the use of Hall charts, Nichols charts, inverse Nyquist plots and root locus plots. The graphical methods also indicate the relative stability of a system. Numerous commercial computer packages are available (see, for example, references 3 to 5) to assist the designer of control systems. These packages include the usual graphical representations and can be obtained from the suppliers whose addresses are given in the references. The reader is also referred to the Bibliography given at the end of this chapter for a more comprehensive coverage of these methods and techniques.

11.1.4.3 Effect of trasnport delay

The influence of a transport, or time, delay on the response of an underdamped second-order system to a step input is displayed in *Figure 11.24*. Although it is virtually impossible to account for a time delay in a differential equation, it is simply accommodated in the frequency-domain model as an additional element in the system block diagram. In the frequency-domain model, the time delay effects a phase shift of $-\omega T$ and can be expressed as

$$\text{Time delay} = e^{-sT} = e^{-j\omega T} \quad (11.30)$$

Consider the open-loop response of a first-order system incorporating a time delay as illustrated in *Figure 11.25*. The open-loop transfer function becomes

$$\frac{Y}{X} = \frac{Y(s)}{X(s)} = G(s) = \frac{ke^{-sT}}{1 + \tau s} \quad (11.31)$$

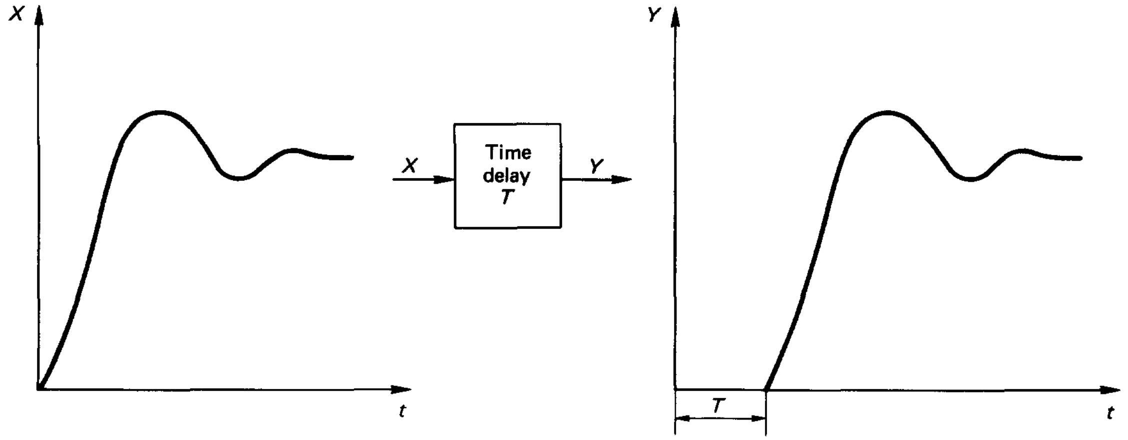


Figure 11.24 Effect of transport delay

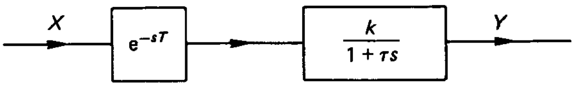


Figure 11.25 First-order system with a time delay

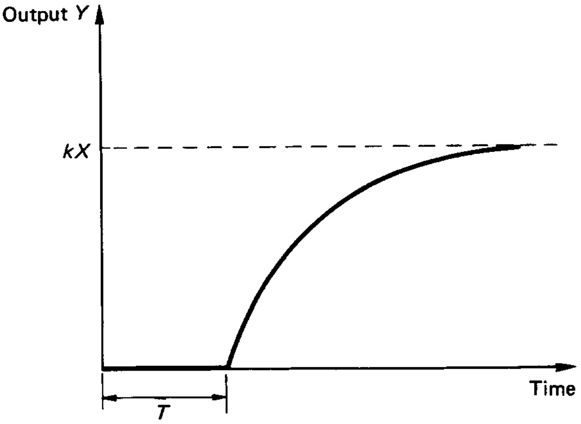


Figure 11.26 First-order, open-loop system response with a time delay

$$\text{Amplitude ratio} = \left|\frac{Y}{X}\right| = \frac{k}{\sqrt{1+\omega^2\tau^2}} \quad (11.32)$$

$$\text{Phase lag} = \phi = -[\omega T + \tan^{-1}(\omega\tau)] \quad (11.33)$$

The first-order, open-loop system response with a time delay is shown in *Figure 11.26*. Similarly, a time delay in a system with feedback does not alter the amplitude ratio but adds a phase shift to the frequency response. The system therefore becomes less stable and, in some cases, it may be necessary to reduce the closed-loop gain in order to obtain a stable response. The penalty to be paid for increasing the stability in this manner is an increase in the steady-state error.

11.2 Control strategies

The basic closed-loop system with common symbol representation is shown in *Figure 11.27*. The nomenclature used in the figure is defined as follows:

SP(s) is the set point (the required value $r(t)$ is sometimes used);
PV(s) is the process value (the corrected value $c(t)$ is sometimes used);
$E(s)$ is the error signal, which is the difference between SP and PV;
$U(s)$ is the control effort output from the controller to the process;
$C(s)$ is the controller transfer function; and
$G(s)$ is the process transfer function.

The transfer function for the closed-loop system is obtained as before:

$$\text{PV}(s) = C(s)G(s)E(s) \quad (11.34)$$
$$= C(s)G(s)[\text{SP}(s) - \text{PV}(s)]$$

Hence,

$$\frac{\text{PV}(s)}{\text{SP}(s)} = \frac{C(s)G(s)}{1 + C(s)G(s)} \quad (11.35)$$

11.2.1 ON/OFF control

In many applications, a simple ON/OFF strategy is perfectly adequate to control the output variable within preset limits. The ON/OFF control action results in either full or zero power being applied to the process under control. A mechanical type of thermostat proides a good example of an ON/OFF based controller. The ON/OFF control strategy results in an output which fluctuates about the set point, as illustrated in *Figure 11.28*.

ON/OFF controllers usually incorporate a 'deadband' over which no control action is applied. The deadband is necessary to limit the frequency of switching between the ON and OFF states. For example, in a temperature control system, the ON/OFF control strategy would be:

(1) if temperature $< T_{min}$, then heater is to be switched ON;
(2) if temperature $> T_{max}$, then heater is to be switched OFF.

The deadband in the above case is $(T_{max} - T_{min})$, and while the temperature remains within the deadband no switching will occur. A large deadband will result in a correspondingly large fluctuation in the process value about the set point. Reducing the deadband will reduce the level of fluctuation but will increase the frequency of switching. The simple ON/OFF control strategy is mostly applicable to processes and systems which have long time constants and, consequently, have relatively slow response times, e.g. temperature and level control.

While being simple in concept, ON/OFF control systems are, in fact, highly non-linear and they require some complex non-linear technique to investigate their stability characteristics.

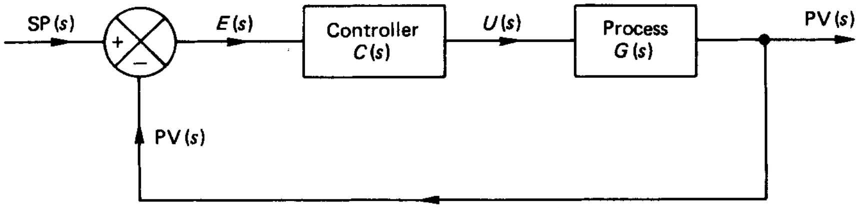


Figure 11.27 Basic closed-loop control system

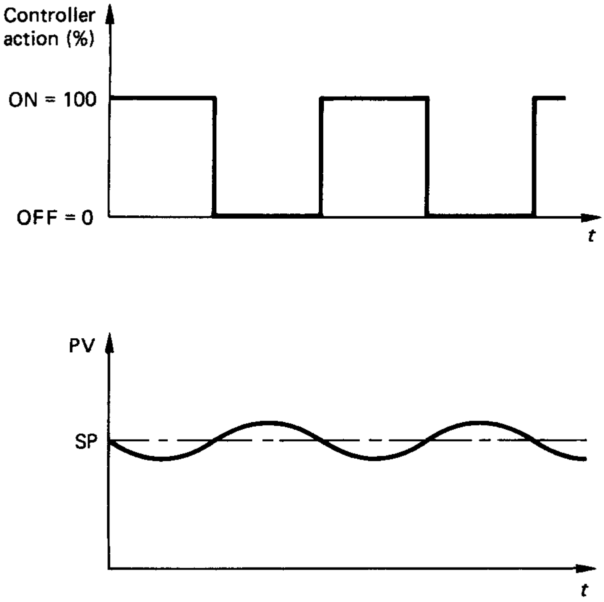


Figure 11.28 Output variation with ON/OFF control

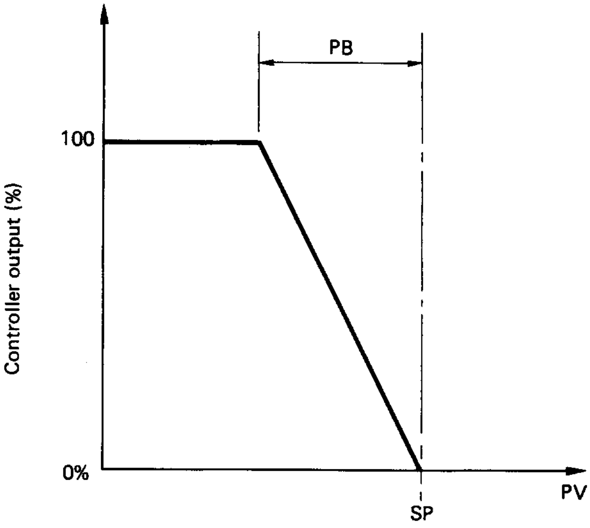


Figure 11.29 Illustration of the proportional band

11.2.2 Three-term, or PID, control

Since complicated transfer functions can be very difficult to model, the most common strategy used to define the controller transfer function is the so-called 'three term', or PID, controller. PID is the popular acronym for 'proportional, integral and derivative'. The three elements of the controller action U based on the evaluated error E are described below.

11.2.2.1 Proportional action

$$\text{Controller output} = KE \tag{11.36}$$

where K is the controller gain.

Manufacturers of three-term controllers tend to favour the parameter 'proportional band' (PB) in preference to gain K. The proportional band represents the range of the input over which the output is proportional to the input. The PB is usually expressed as a percentage of the input normalised between 0 and 100% (see *Figure 11.29*).

To illustrate the concept of proportional band, a temperature control application can be considered where the set point is, say, 80°C and the proportional band is set to, say, 5% over a measured temperature span of 0–100°C. The actual proportional band is, therefore, 5°C and proportional action will apply over the temperature range 75–80°C. If the temperature is below 75°C, then 100% of the available power will be supplied to the heating device. Between 75°C and 80°C, a proportion of the available power will be applied to the heating device, as shown in *Figure 11.29*. For temperatures in excess of 80°C, 0% of the available power is supplied. It should be apparent that 'proportional band' is a more meaningful term than 'gain'. The two parameters, however, are very simply related:

$$\text{PB\%} = 100/K \tag{11.37}$$

It is also apparent from *Figure 11.29* that, as the proportional band is decreased, then the control action tends towards an ON/OFF strategy. A very large proportional band will result in a somewhat sluggish response.

It must also be noted that, for proportional control only, there must always be an error in order to produce a control action. From equation (11.35), proportional control only gives a transfer function of the form:

$$\frac{\text{PV}(s)}{\text{SP}(s)} = \frac{KG(s)}{1 + KG(s)} = \frac{1}{(1/KG(s)) + 1} \tag{11.38}$$

For steady-state conditions, s tends to 0 and $G(s)$ tends to a constant value. Equation (11.38) shows, therefore, that the gain must theoretically tend to infinity if PV = SP, and the steady-state error is to approach zero.

This is simply another manifestation of the classical control problem; i.e. stability at the expense of accuracy, and vice versa. With a very high gain, i.e. a low proportional band, the steady-state error can be very much reduced. A low proportional band, however, tends to ON/OFF control action, and in sensitive systems a violent unstable oscillation may result.

11.2.2.2 Integral action

The limitations of proportional control can be partly alleviated by adding a controller action which gives an output contribution that is related to the integral of the error value with respect to time:

$$\text{Controller output} = K_i \int E\,dt \tag{11.39}$$

where K_i is the controller integral gain. $K_i = K/T_i$ and T_i is the controller 'integral time', or 'reset'.

The nature of integral action (equation (11.39)) suggests that the controller output will increase monotonically as long as an error exists. As the error tends to zero the controller output tends towards a steady value. The general behaviour of the controller output with integral action is shown in *Figure 11.30*.

If T_i is very large, the integral action contribution will be low and the error may persist for a considerable time. If, on

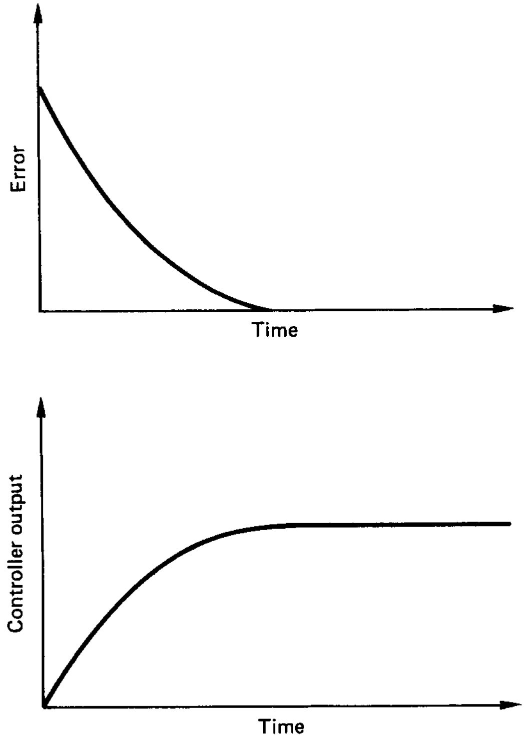


Figure 11.30 Controller output with integral action

the other hand, T_i is too small, the magnitude of the integral term may cause excessive overshoot in the output response. Unstable operation is also possible when T_i is too small and the controller output value then increases continuously with time.

11.2.2.3 Derivative action

The stability of a system can be improved and any tendency to overshoot reduced by adding derivative action. Derivative action is based on the rate of change of the error, i.e.

$$\text{Controller output} = K_d \frac{dE}{dt} \tag{11.40}$$

where $K_d = KT_d$ is the controller derivative gain, and T_d is the controller 'derivative time', or 'rate'.

Equation (11.40) indicates that the derivative action is dependent on how quickly, or otherwise, the error changes. Derivative action tends, therefore, only to come into operation during the early transient part of the response of a system.

The full three-term control strategy may be written as:

$$K\left[E + \frac{1}{T_i}\int E\,dt + T_d \frac{dE}{dt}\right] \tag{11.41}$$

To summarise, the proportional action governs the speed of the response, the integral action improves the accuracy of the final steady state, and the derivative action improves the stability. Note that derivative action may result in poor performance of the system if the error signal is particularly noisy.

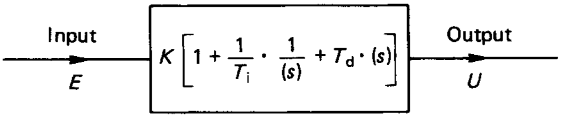


Figure 11.31 Three-term, or PID, control

In Laplace notation, the three-term controller transfer function is as shown in *Figure 11.31*.

11.2.3 Empirical rules for PID controller settings

A simple, and still popular, technique for obtaining the controller settings to produce a stable control condition is that proposed by Ziegler and Nichols.[6] The technique is purely empirical and is based on existing, or measurable operating records of the system to be controlled.

11.2.3.1 Open-loop 'reaction curve' method

The process to be controlled is subjected to a step input excitation and the system open-loop response is measured. A typical open-loop response curve is shown in *Figure 11.32*.

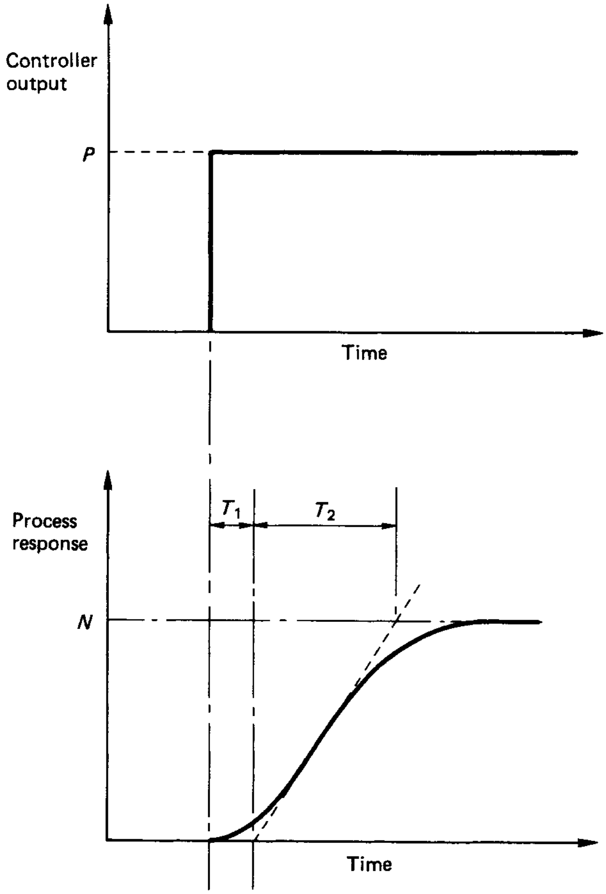


Figure 11.32 Open-loop system response to a step input

Any system which has a response similar to that shown in *Figure 11.32* has a transfer function which approximates to a first-order system with a time delay, i.e.

$$G(s) = \frac{ke^{-sT_1}}{1 + T_2 s} \qquad (11.42)$$

In general industrial applications, oscillatory open-loop responses are extremely rare and *Figure 11.32* is in fact representative of quite a large number of real practical processes.

If N is the process steady-state value for a controller step output of P, the system steady-state gain is

$$k = N/P \qquad (11.43)$$

The 'apparent dead time' T_1 and the 'apparent time constant' T_2 can be measured directly from the process response curve. The three parameters k, T_1 and T_2 are then used in a set of empirical rules to estimate the optimum controller settings. The recommended controller settings are given in *Table 11.1*.

In fast-acting servomechanisms, where T_1 may be very small, the method is not very successful. For moderate response systems, however, the method will yield very reasonable first approximation controller settings.

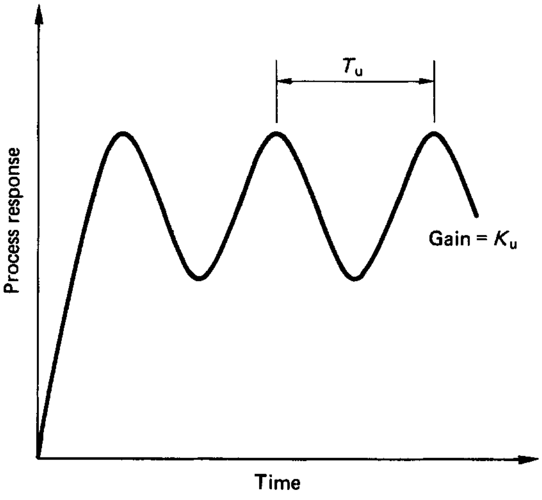


Figure 11.33 Continuous cycling method

11.2.3.2 Closed-loop 'continuous cycling' method

The process to be controlled is connected to the PID controller and the integral and derivative terms are eliminated by setting $T_d = 0$ and $T_i = \infty$. In some industrial controllers the integral term is eliminated, with $T_i = 0$. A step change is introduced and the system run with a small controller gain value K. The gain is gradually increased for a step input until constant-amplitude oscillations are obtained, as illustrated in *Figure 11.33*.

The gain K_u, which produces the constant amplitude condition, is noted and the period of the oscillation T_u is measured. These two values are then used to estimate the optimum controller settings according to the empirical rules listed in *Table 11.2*.

For a temperature control system, typical values of T_u are about 10 s for a tungsten filament lamp, 2 min for a 25 W soldering iron and 10–30 min for a 3 kW heat treatment furnace.

The PID settings obtained according to the Ziegler and Nichols methods are approximate only, and some 'fine tuning' would almost certainly be required in practice.

11.2.4 Three-term controller with a first-order system

The block diagram of the system is depicted in *Figure 11.27* and equation (11.35) defines the closed-loop transfer function.

Table 11.1 Optimum controller settings according to Ziegler and Nichols[6]

Control action	*K*	T_i	T_d
P	$T_2/(T_1 k)$	—	—
P + I	$(0.9\ T_2)/(T_1 k)$	$T_1/0.3$	—
P + I + D	$(1.2\ T_2)/(T_1 k)$	$2T_1$	$0.5\ T_1$

Table 11.2 Optimum controller settings according to Ziegler and Nichols[6]

Control action	*K*	T_i	T_d
P	$0.5\ K_u$	—	—
P + I	$0.45\ K_u$	$T_u/1.2$	—
P + I + D	$0.6\ K_u$	$T_u/2$	$T_u/8$

If a P + I controller is to be used, i.e. no derivative action, then the controller transfer function is

$$C(s) = K\left[1 + \frac{1}{T_i s}\right] \qquad (11.44)$$

The process is modelled as a first-order system and its open-loop transfer function is given by equation (11.14).

Substituting equations (11.14) and (11.44) into equation (11.35) results, after some manipulation, in

$$\frac{\mathrm{PV}(s)}{\mathrm{SP}(s)} = \frac{\frac{kK}{\tau T_i}(1 + sT_i)}{\left(s^2 + \left(\frac{1}{\tau} + \frac{kK}{\tau}\right)s + \frac{kK}{\tau T_i}\right)} \qquad (11.45)$$

Comparing the denominator with that for the generalised second-order system, i.e. equation (11.21), it can be shown that

$$2\zeta\omega_n = \left(\frac{1}{\tau} + \frac{kK}{\tau}\right) \qquad (11.46)$$

and

$$\omega_n^2 = (kK/T_i\tau) \qquad (11.47)$$

For the system being controlled, both k and τ are known either via a mathematical model, or an open-loop test. The controller settings K and T_i can then be calculated for a chosen damping ratio ζ and natural frequency ω_n. Alternatively, a

controller gain can be imposed and the corresponding natural frequency evaluated.

For full PID control, an initial value of $T_d = T_i/4$ can be used. Other systems can be similarly handled to obtain the approximate PID controller settings. In all cases some fine adjustment would probably be necessary to obtain the optimum output response.

11.2.5 Disturbance sensitivity

The main problem with the classical single-loop control system is that it is not truly representative of the natural environment in which the system operates. In an ideal single-loop control system the controlled output is a function only of the input. In most practical systems, however, the control loop is merely a part of a larger system and is therefore subject to the constraints and vagaries of the larger system. This larger system, which includes the local ambient, can be a major source of disturbing influences on the controlled variable. The disturbance may be regarded as an additional input signal to the control system. Therefore, any technique which is designed to counter the effect of the disturbance must be based on a knowledge of the time-dependent nature of the disturbance and also its point of entry into the control system. Two methods commonly used to reduce the effect of external disturbances are 'feedforward' and 'cascade' control.

11.2.5.1 Feedforward control

The principle of a feedback loop is that the output is compared with the desired input and a resultant error signal acted upon by the controller to alter the output as required. This is a control action which is implemented 'after the fact'. In other words, the corrective measures are taken after the external disturbance has influenced the output. An alternative control strategy is to use a feedforward system where the disturbance is measured. If the effect of the disturbance on the output is known, then theoretically the corrective action can be taken before the disturbance can significantly influence the output. Feedforward can be a practical solution if the external disturbances are few and can be quantified and measured. A block diagram illustrating the feedforward concept is shown in *Figure 11.34*.

Feedforward control can be difficult to implement if there are too many, or perhaps unexpected, external disturbances. In *Figure 11.34*, the path which provides the corrective signal appears to go back. The strategy is still feedforward, however, because it is the disturbance which is measured and the corrective action which is taken is based on the disturbance and not the output signal. Some control systems can be optimised by using a combination of feedforward and feedback control.

11.2.5.2 Cascade control

Cascade control is implemented with the inclusion of a second feedback loop and a second controller embodied within a main feedback loop in a control system (see *Figure 11.35*). The second feedback loop is only possible in practice if there is an intermediate variable which can be measured within the overall process. Cascade control generally gives an improvement over single-loop control in coping with disturbance inputs. The time constant for the inner loop is less than that for the component it encloses and the undamped natural frequency of the system is increased. The overall effects of cascade control are an increase in the system bandwidth and a reduction in the sensitivity to disturbances entering the inner loop. Disturbances entering the outer loop are unaffected. Cascade control works best when the inner loop has a smaller time constant than the outer loop.

11.2.6 Direct digital control

Most of the standard texts on control engineering are centred on the mathematical modelling of systems and processes, and subsequently on the stability considerations of these entities. This approach requires detailed knowledge of the system constituent parts to enable the formulation of a suitable differential equation to describe the dynamic behaviour. It is often only in the idealised world of servomechanisms that adequate models can be derived. For many real processes an adequate system model can be difficult, if not impossible, to obtain.

The modern emphasis is therefore on the application of computer based control strategies which can be made to work with real systems. The recent developments in microelectronics, particularly microprocessors, has made microcomputer devices the natural choice as the controller for many systems. The microcomputer allows the implementation of such functions as arithmetic and logic manipulation, timing and counting. With many analogue input/output modules available to interface to the microcomputer, the overall 'intelligence' of the system is greatly enhanced.

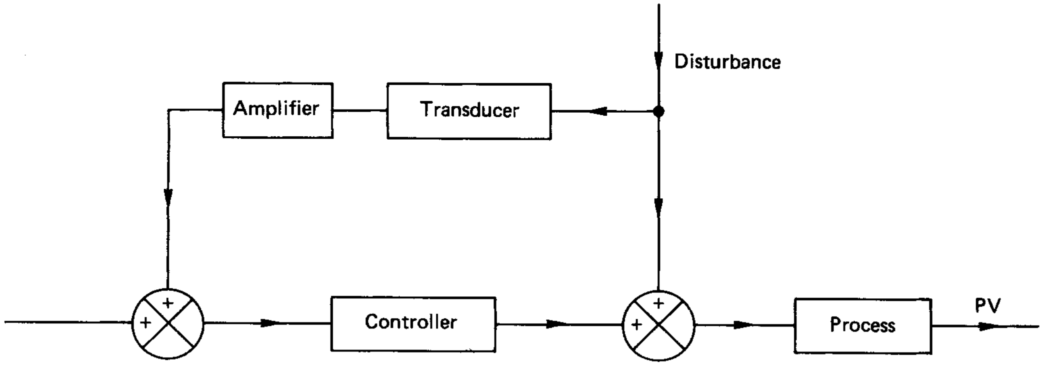


Figure 11.34 Feedforward control system

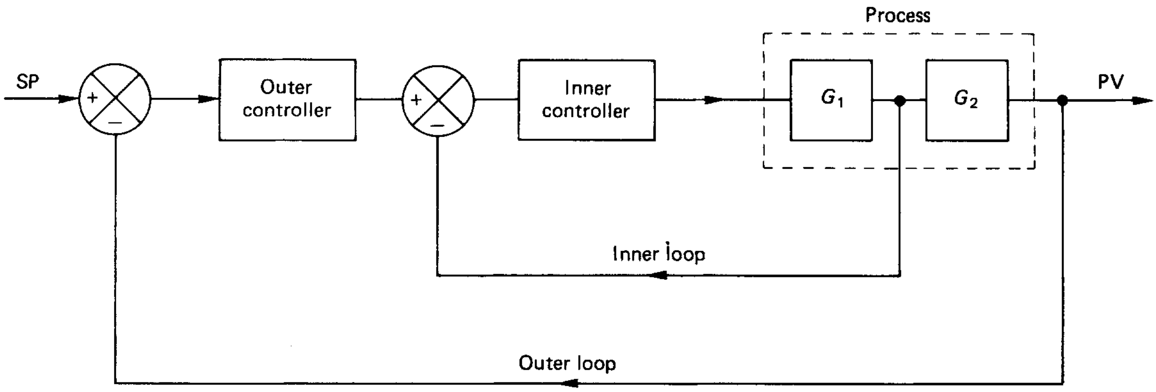


Figure 11.35 Cascade control system

The basic elements of the computer based control system include the microprocessor, memory, an input interface to measure the process variable and an output interface to supply power to the controlled process. The control effort output to the process is determined by the control strategy which takes the form of an algorithm incorporated within the computer software. The fundamental digital based control system is depicted in *Figure 11.36*.

In the generalised layout given in *Figure 11.36*, the microcomputer performs a number of tasks which would require separate elements in an equivalent analogue system. The two inputs to the microcomputer are the desired set point and a signal from the process via a feedback loop. The term 'process' is being used in a quite arbitrary sense in this context. The microcomputer firstly performs the function of comparing the process value with the set point to establish the error. The control strategy is then applied to the error value to determine the corrective action necessary. The microcomputer subsequently outputs the appropriate signal to the process via other additional elements in the system. These additional elements include the input–output interfaces between the digital-based computer and the otherwise analogue-based control system. Digital to analogue converters (DACs) and analogue to digital converters (ADCs) are featured in Section 11.5. The transducer, which provides the link between the physical world and the electronically based control system, is covered in its various forms in Section 11.3. The essential fundamentals of microprocessor technology are outlined in Section 11.4, and the applications of microcomputer based control are described in Section 11.6.

11.2.7 Adaptive and self-tuning control

The concept of adaptive control is based on the ability to measure the system behaviour at any time and to alter the controller settings automatically to provide an optimum system response. Adaptive control has been a very active research topic over the last 10 years, but it is only recently that practical applications using adaptive controllers have appeared.

The simplest approach to adaptive control is the so-called 'gain scheduling' method (*Figure 11.37*). The principle of gain

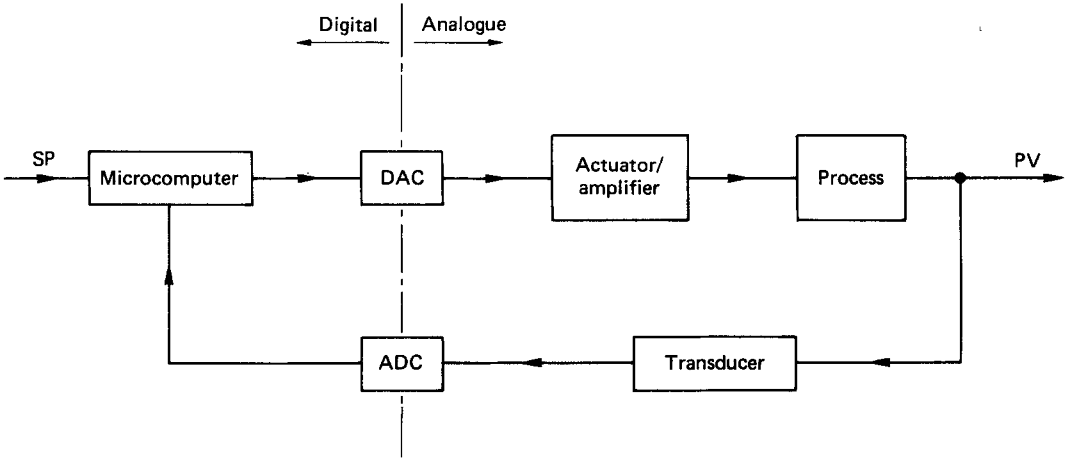


Figure 11.36 Fundamental digital-based control system

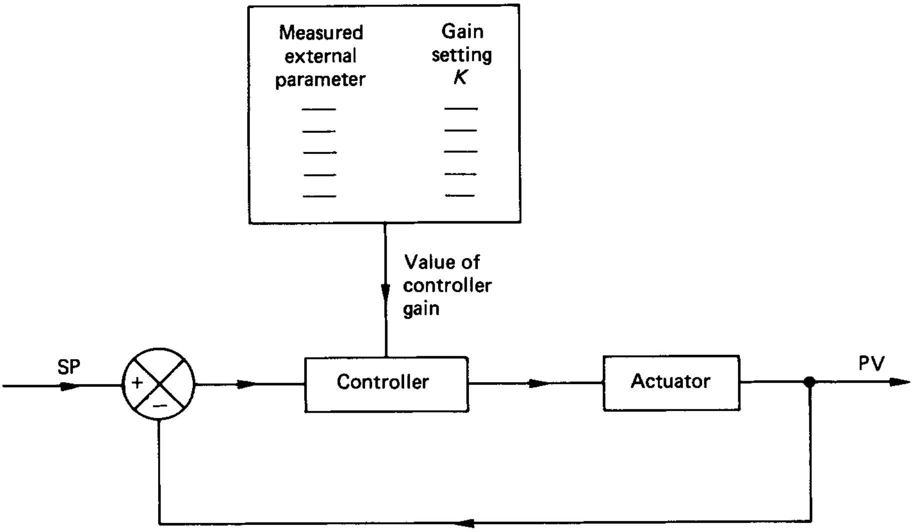


Figure 11.37 Adaptive control by gain scheduling

scheduling is that some relevant external parameter is measured and an appropriate value of gain is then selected for the controller. Gain scheduling was first developed for aileron control in high-altitude aircraft. The low air density at high altitude has a profound effect on the in-flight dynamics and the purpose of gain scheduling was to provide the pilot with a more consistent 'feel' of the aircraft's handling, independent of altitude. Gain scheduling has the advantage that the system stability margins can be well established for any value of gain and the technique is generally fast acting. The method is limited, however, since the gain adjustment is a function of only one measured parameter. In most systems the process may be subject to any number of external parameters, and the more modern adaptive controllers use some mathematical model as a basis of comparison with the actual control system (*Figure 11.38*).

The mathematical model shown in *Figure 11.38* receives the same input as the actual system and an error is created relating the difference between the actual and the model system output. The error may then be used as a basis for altering the controller settings. Obviously, the quality of the control will

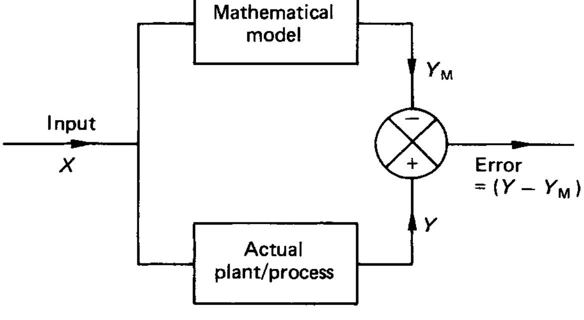


Figure 11.38 Model/actual error generation

depend on how well the model reflects the actual system. The usual implementation of model reference adaptive control is illustrated in *Figure 11.39*.

It is worth noting that the original feedback loop is left intact such that failure of the adaptive loop will not render the system inoperative. External disturbances operating on the actual plant will change the actual/model error signal and provide the basis for retuning the controller settings via the adaptive loop. The adjustment of the controller settings implies that there must be some well-defined strategy to determine the level and nature of the adjustments made.

Self-tuning control takes the adaptive concept one stage further, in that the mathematical model of the system is also updated as more input and output data from the actual system are acquired. A schematic diagram of a self-tuning controller is shown in *Figure 11.40*. The computer based self-tuning controller estimates the system dynamics and then uses this estimate to implement the optimum controller settings. The continuous updating of the system parameters at each sampling interval is called 'recursive parameter estimation'. Previously estimated parameters are also available, and these can be used in perhaps a 'least-squares' method to provide some overall smoothing of the control function. With the latest system parameters available, the self-tuning controller then goes through some 'design' procedure to optimise the controller settings. This 'design' is usually based on the desired output response of the system. One particular design procedure is based on the root locus method for stability analysis. By adjustment of gains and time constants in the control algorithm, the method seeks to tune the transfer function and thereby govern the output response. Other procedures are often based on the rules of Ziegler and Nichols.[6] The final process in the self-tuning control cycle is the physical imposition of the optimised controller settings on the actual system.

Self-tuning control is generally applied to the more complex processes where transportation delays, non-linearities and multiple control loops greatly add to the complexity. The stability of such systems is in most cases non-deterministic since there is no generalised theory available. Traditionally then, most self-tuning controllers are based on well-

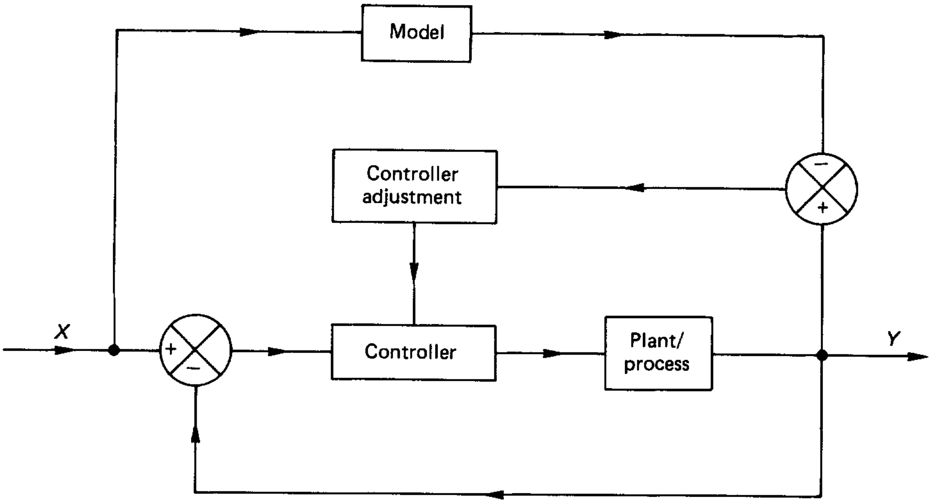


Figure 11.39 Model reference adaptive control

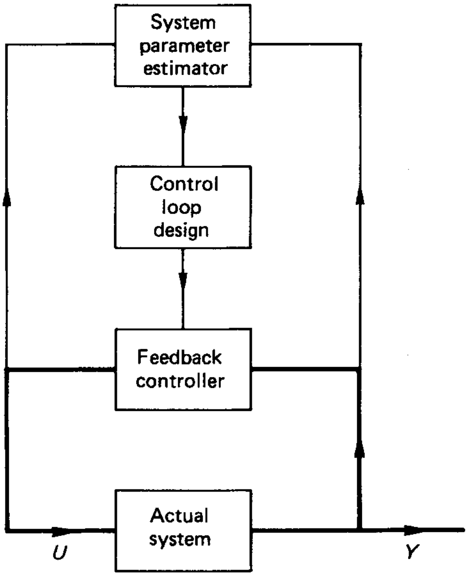


Figure 11.40 Self-tuning controller

established three-term control principles, but with the added enhancement of adaptability. A number of proprietary self-tuning controllers are available commercially (e.g. Kraus and Myron[7] describe the Foxboro Company's 'EXACT' controller). The EXACT controller is based on PID principles and uses the Ziegler and Nichols[6] rules in the self-tuning mode.

11.2.8 Sampled-data systems

The two previous subsections gave an overview of direct digital control and the natural progression to adaptive and self-tuning controllers. The common factor which relates these concepts is the use of a computer, or microcomputer, as a central feature of the control system. The computer acts as the compensator in the control loop and the analogue-to-digital and digital-to-analogue interfaces provide the link between the digital-based computer and the otherwise analogue-based controlled system. Being digitally based, the computer operates in discrete time intervals and, indeed, the control strategy, which exists in the software, must also take a finite time for its evaluation and implementation. Time delays are also inevitable in the analogue-to-digital and the digital-to-analogue conversion processes and these cumulative time delays result in what is called a 'sampled-data system'. The difference between a sampled datum, or a discrete signal, and its continuous counterpart is displayed in *Figure 11.41*. The closure time q is the time taken to complete the digitisation of the instantaneous signal. Generally, $q \ll T$.

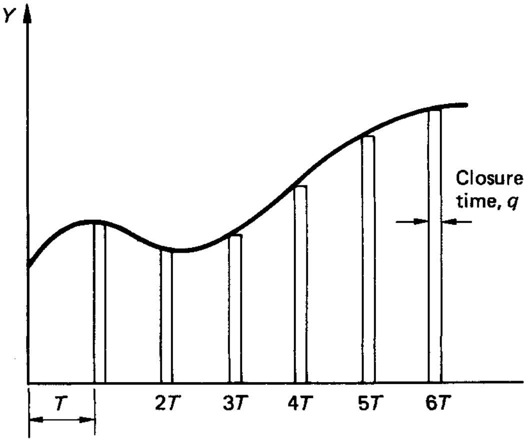


Figure 11.41 Digitisation of a continuous signal

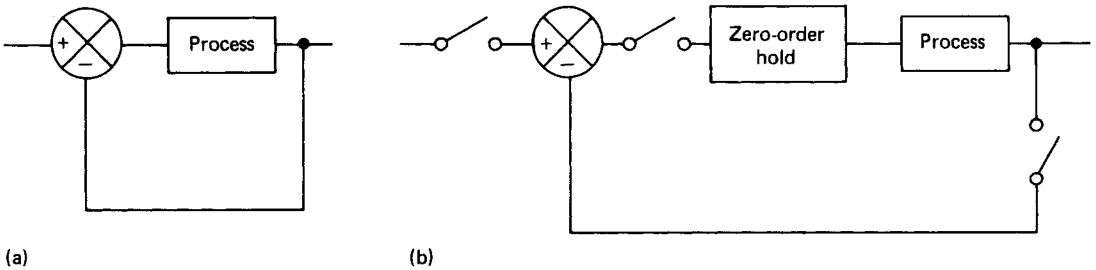


Figure 11.42 Continuous (a) and discrete (b) closed-loop control systems

It is apparent that much less information is available in the sampled-data signal as it exists only as a pulse train, interspersed with gaps in the information between the sampled points. If the sampling frequency is high enough, then this need not be troublesome (see also Section 11.6.4). The inevitable additional time delays in a sampled-data system, however, have implications regarding the overall stability of the system.

The Laplace transform method cannot be used to analyse a sampled-data system, but there is a related transform which is applicable to discrete time systems known as the 'z-transform'. The relation is:

$$z = e^{sT} \tag{11.48}$$

The symbol z is associated with a time shift in a difference equation in the same way that s is associated with differentiation in a differential equation. Equation (11.48) then gives a conformal mapping from the s plane to the z plane and provides the means for analysing discrete time systems. The general method of solution involves the derivation of the closed-loop transfer function in terms of the Laplace variable. The equivalent discrete time system is then represented by introducing a 'zero-order hold' to account for the additional time delays in the discrete system (*Figure 11.42*).

The transfer function for a zero-order hold is:

$$\frac{1 - e^{-sT}}{s} \tag{11.49}$$

The zero-order hold is simply included in the evaluation of the closed-loop transfer function for the discrete-time system. The next step is to replace the Laplace transforms with their equivalent z transforms. The resulting transfer function in terms of z transforms can then be analysed for stability in much the same manner as the root-locus method is used for continuous systems. A comprehensive coverage of z transform techniques and their application to the stability analysis of sampled-data systems is given by Leigh.[8]

11.2.9 Hierarchical control systems

The ultimate aim in industrial optimisation is the efficient control of complex interactive systems. Recent hardware developments and microprocessor-based controllers with extensive data-handling power and enhanced communications have opened up possibilities for the control of interlinked systems. What is required, but not yet realised, is a theoretical framework on which to base the analysis of such systems. Nonetheless, and in the absence of theory, hierarchical control systems do exist and are currently being used effectively in the control of various large-scale plant and processes.

The usual approach adopted is to subdivide the complex system into a number of more manageable parts. This is the concept of hierarchical control which might be thought of as a subdivision in decreasing order of importance. Hierarchical control exists in two basic forms: multilayer and multilevel. Multilayer control is that in which the control tasks are subdivided in order of complexity. Multilevel control, on the other hand, is that where local control tasks are coordinated by an upper echelon of supervisory controllers. Multilayer control is illustrated concisely in an elaborate adaptive-type controller. The hierarchy is depicted in *Figure 11.43*.

The first level is that of regulation which is characterised by the classical single-closed-loop control system. Moving up the hierarchy we have optimisation of the controller parameters. Optimisation is representative of the basic adaptive controller, using simple gain scheduling or some model reference criterion. The next highest level is that of parameter adaptation. Parameter adaptation is embodied in the self-tuning controller which represents the beginnings of an 'expert system' approach. The highest level is that of model adaptation which is based on long-term comparisons between the model and the actual performance. If the system is modelled accurately to

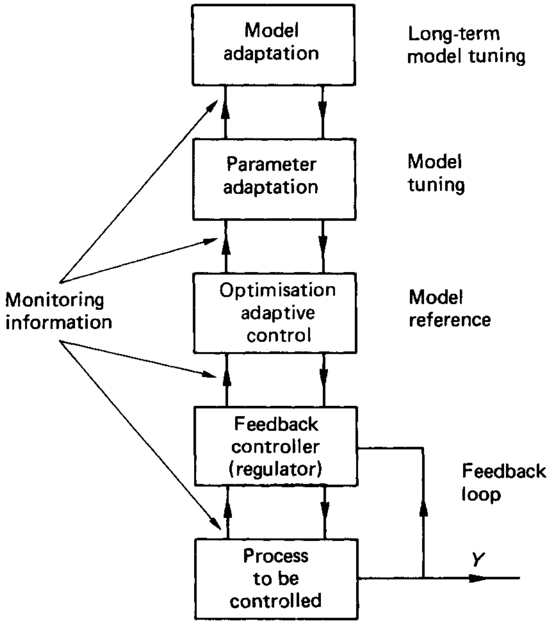


Figure 11.43 Multilayer control system

begin with, the model adaptation level might only rarely be entered.

Multilevel control is characterised as local controllers whose actions are governed by higher levels of supervisory controllers. The local controllers operate independently to achieve local targets. The function of the supervisory controller is to reconcile the interaction of the local controllers to achieve the 'best' overall performance. the multilevel concept has some similarity with cascade control but is not so amenable to analysis. Multilevel control gives rise to a pyramid-like structure, typified by that shown in *Figure 11.44*.

At the base of the pyramid are the local controllers, monitoring and adjusting individual parameters in the overall process. At the next highest level the supervisory controllers 'oversee' a more complete picture of the process. The intermediate supervisory controllers have more input data to contend with, and they might perhaps relax the control on one of the process variables while tightening up on another. This, of course, would only be done to benefit the process overall. The highest level of supervisory controller has the responsibility for the entire process. This controller may have access to additional input data which are not available to any of the lower level controllers. The main supervisory controller then is in overall 'command' and can influence any of the 'subordinate' controllers. The similarity between multilevel control and the organisation structure of an industrial company is not merely coincidental. The latter is the structural model upon which the former is based.

11.3 Instrumentation and measurement

The implementation of any control action is based on the error which exists between the process value and the required set point. Establishment of the error requires a measurement to be made of the process value. In some cases, the set point is also a measured parameter. For an electronically based control system, either analogue or digital, the connection between the real physical world and the control system is transmitted by some form or other of transducer.

11.3.1 Analogue transducers

The function of a transducer is, in the broadest sense, one of energy conversion. To this end an electric motor could be considered as a transducer which converts electrical power into mechanical power. In the context of instrumentation and measurement, the transducer provides the means of conversion between some physical variable, which is representative of the process, to a proportional voltage or current. The proportional voltage, or current, then forms the basic signal upon which the control of the process is based.

In manufacturing engineering, the most commonly measured parameters, for control purposes, are position and speed. These two parameters form the basic feedback variables associated with the control of machine tools, robots and automatic guided vehicles. Feedback control may also be based on a measurement of force, or on some visual criterion, particularly in the control of robotic arms and manipulators. In the general field of process control, the measured variables might also include temperature, flow rate, torque, power, displacement, angle, strain, humidity, light intensity, acidity, sound or perhaps simply time. The list as given probably represents about 90% of all the measurements made in the industrial environment. With the appropriate transducer and signal processing, all of the above parameters can be converted into a proportional output voltage.

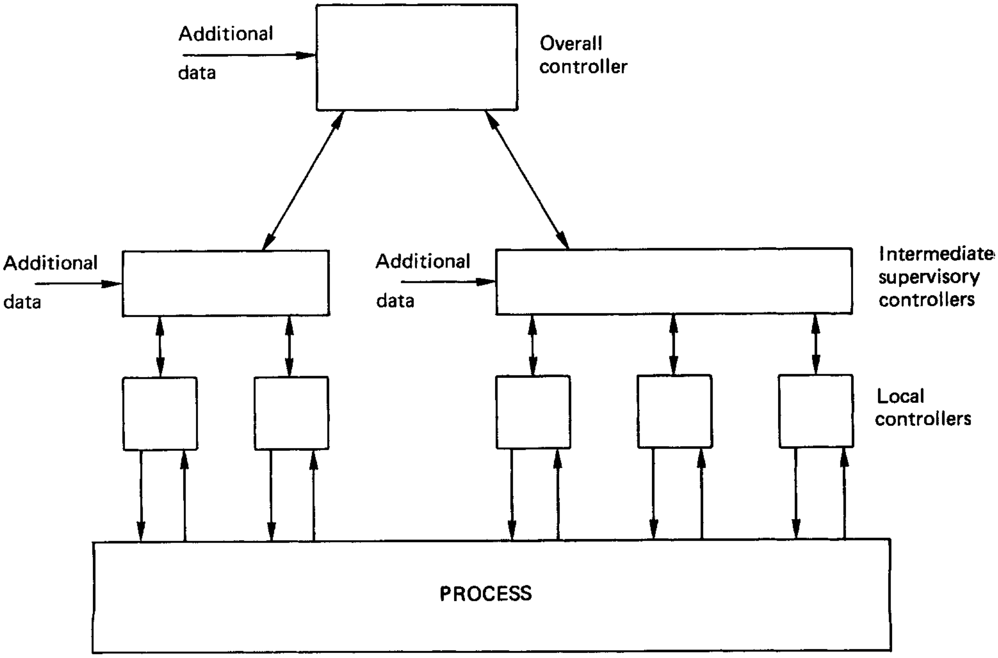


Figure 11.44 Multilevel control system

11.3.2 Digital transducers

Most conventional transducers have an analogue output voltage which could be measured using fairly traditional instrumentation, including for example an oscilloscope or a voltmeter. The output of many transducers, however, is in the millivolt range and, more often than not, the basic output must be further processed, or conditioned, to produce a useful signal. The processing required might only involve simple amplification but, if the controller is digitally based, then additional signal processing is needed to convert the analogue voltage into a digital representation. This latter processing is performed using an analogue-to-digital converter (see Section 11.5.3). An obvious advantage would be a transducer which can produce a direct digital output. Such a device could then dispense with the intricacies both of signal conditioning and analogue to digital conversion. There are, however, very few devices available which can easily produce a digital output in response to a physical variable. The absolute digital shaft encoder (*Figure 11.45*) is one of the few transducers which satisfy the digital-output specification in the strictest sense.

The device incorporates a sectioned disc which is mounted on a rotating element. The disc is divided into a number of tracks, three in this case, and a number of sectors. If the number of tracks is denoted by m, then the number of sectors is numerically given by 2^m. With $m = 3$, the number of sectors is 8 and this also denotes the number of angular resolvable sections on the disc. For each resolvable section a unique pattern is generated. These patterns may be detected by optical, magnetic or direct contact methods. Resolution may be improved with the addition of more tracks. Twelve tracks, for example, give an angular resolution to within ±0.09° rotation. The absolute shaft encoder may be used to measure either rotational speed or angular displacement.

In *Figure 11.45* a simple binary code is used to illustrate the principle of the absolute shaft encoder. It can be seen that between sectors 3 and 4 and sectors 7 and 0, all the tracks undergo a change in pattern. When all the tracks change there must always be one which changes first before the others. Any misalignment of the track sensors, therefore, could generate a considerable error in rotational-position measurement. For this reason, practical shaft encoders use some 'unit distance code' in preference to a binary code. A unit distance code, such as the 'Gray code' described in Section 11.4.3, is one where only one track changes pattern for each consecutive resolvable position.

11.3.3 Measurement of angular position

The absolute shaft encoder, discussed above, is a suitable transducer for the measurement of angular position and has the added advantage of a digital output which can be interfaced directly to a computer. A close relative, but one which operates on a rather different principle, is the incremental shaft encoder. The incremental shaft encoder consists of a disc in which a series of holes is arranged around the periphery. A light source, usually a light emitting diode (LED), is positioned on one side of the disc and a photosensitive transistor on the other (*Figure 11.46*).

Rotation of the disc generates a series of light pulses which are registered as a square pulse train output from the photosensitive transistor. By counting the number of pulses relative to a known datum, the angular position of the disc can be determined. By using two light sources and two photosensitive transistors, with a 90° phase difference between the two output signals, the direction of rotation may also be ascertained. The phase relationship between the two signals for clockwise and anti-clockwise rotation is illustrated in *Figure 11.47*. A similar system may equally well be used to measure linear position and direction.

In clockwise, or right-hand movement, the signal from S_2 comes on while the signal from S_1 is still on. For anti-clockwise, or left-hand movement, S_2 comes on while the signal from S_1 is off. This relative switching of the two signals forms the basis of the direction sensing.

In manufacturing applications, high resolution incremental encoders are often used as position feedback sensors, either rotating with a lead screw, or in linear motion with a machine

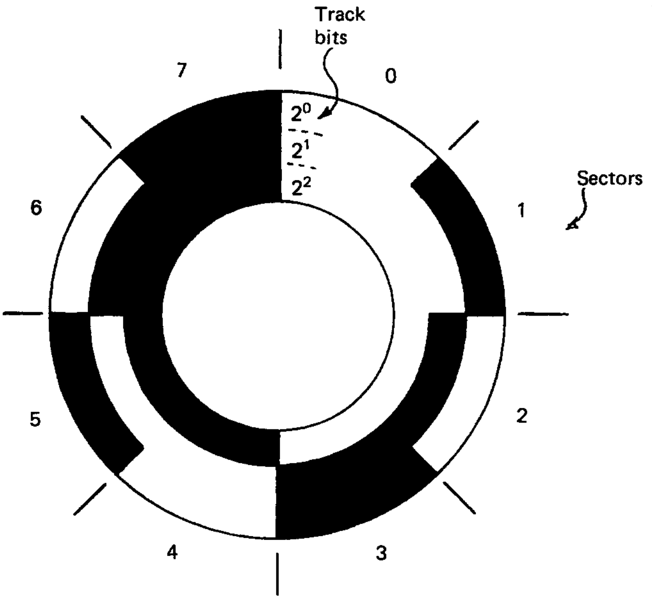


Figure 11.45 Absolute digital shaft encoder (3-bit binary code)

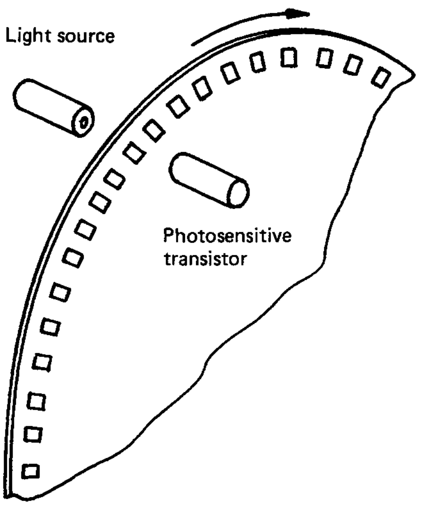


Figure 11.46 Incremental shaft encoder

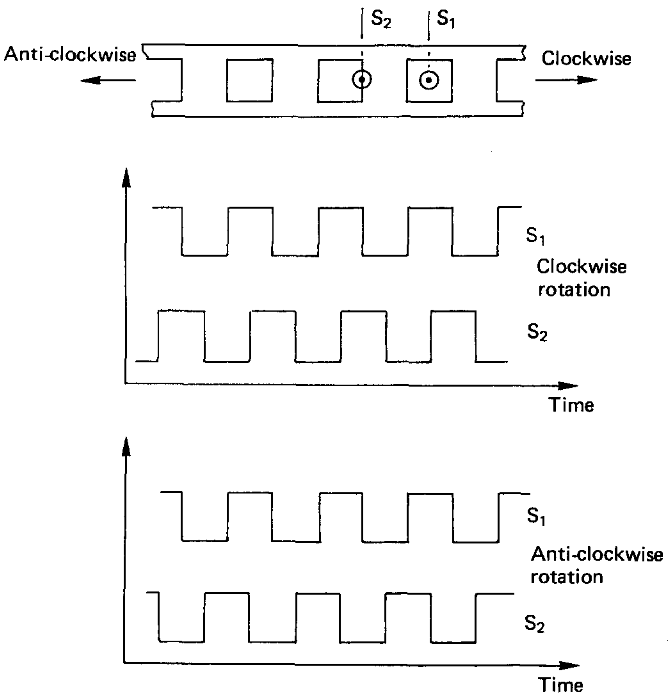


Figure 11.47 Direction sensing with an incremental shaft encoder

slide. They are also used to measure the joint positions in robotic manipulators. In these applications, where high accuracy is essential, the encoders consist of two glass gratings, one of which is fixed and the other which moves. Light which is transmitted through both gratings varies in intensity with the relative displacement of the gratings. The light intensity is picked up by two phototransistors and the signals are used to measure relative position and direction in much the same manner as described previously. The grating pitch is typically 100 per millimetre, but the resolution can be greatly increased by using the moiré fringe effect. Moiré fringes are a diffraction effect which results when the gratings are inclined at a slight angle to each other. The important feature is that the spacing of the interference fringes is many times the spacing of the grid lines. A grid inclination angle of just less than 0.5°, with 100 grid lines per millimetre, would produce a fringe pattern with a 1 mm spacing.

The main disadvantages of the incremental encoder are the possibility of missing a pulse during the count and the loss of the datum reference in the event of power failure. Electrical noise and dirt between the light source and the phototransistor are also problematical. Nonetheless, incremental encoders are cheaper and generally more accurate than absolute encoders.

When accuracy is less important, an angular position transducer based on the variable potential divider (*Figure 11.48*) can be used.

By measuring the voltage between points A and B, a linear relationship between angular displacement, θ and voltage v is obtained:

$$\theta = v(\Theta/V) \qquad (11.50)$$

The maximum voltage occurs when the sliding contact at B reaches the maximum rotation at Θ.

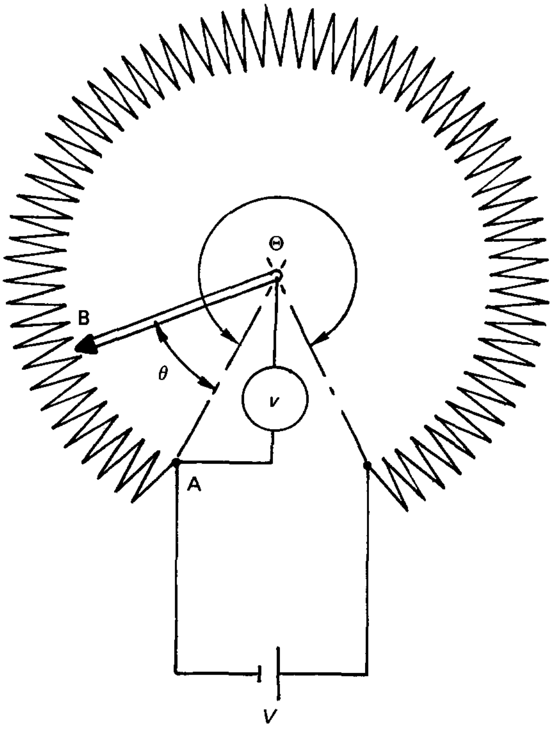


Figure 11.48 Variable potential divider position transducer

As an angular position transducer, however, the variable potential divider has a number of limitations. Equation (11.50) is only strictly applicable if the internal resistance of the voltmeter used to measure v can be considered to be infinite. Taking the internal resistance of the voltmeter R_i into account, Ohm's law for parallel resistors can be used to show that the total resistance between points A and B is:

$$R_{total} = (R_i R_\theta)/(R_i + R_\theta) = R_\theta/([1 + (R_\theta/R_i)] \quad (11.51)$$

where R_θ is the resistance along the arc of resistor element enclosed by the angle θ.

Equation (11.51) shows that R_{total} will approach the value of R_θ only when R_i tends to infinity. This is referred to as a 'loading error' which causes the net resistance between A and B to be reduced due to the additional circuit through the voltmeter.

Further limitations are associated with the sliding contact at B which is a source of friction and inevitable wear.

11.3.4 Measurement of linear displacement

The incremental encoder and the variable potential divider can each be easily adapted for the measurement of linear displacement. The most common transducer used to measure linear displacement however is the linear variable differential transformer (LVDT). LVDTs incorporate an iron core which can move freely within a primary, or power coil and two secondary coils (*Figure 11.49*).

The secondary coils are connected in series opposition and are equally positioned with respect to the primary coil. When the core is centrally located the e.m.f.s generated in the secondary coils are equal and opposite and the net output voltage is zero. Displacement of the core relative to the central zero produces a proportional a.c. output which is converted to a d.c. output via rectification and smoothing circuits. LVDT transducers are essentially immune to friction and wear problems, they have infinite resolution and are highly linear and accurate. LVDTs are available with measurement ranges up to about 1 m.

There are, in addition, a wide range of other displacement transducers which rely on inductive or capacitive effects. These are mostly restricted to the measurement of relatively small displacements of the order of 0.1 mm.

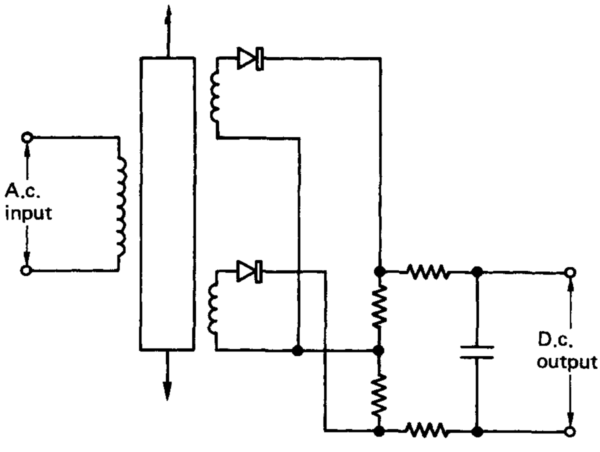


Figure 11.49 Linear variable differential transformer

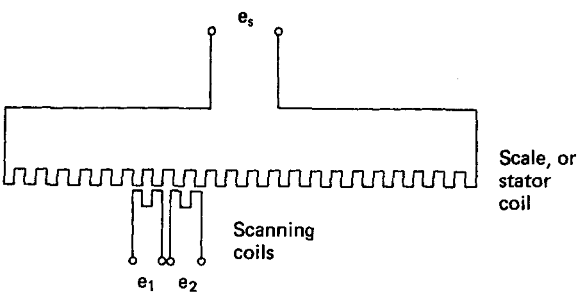


Figure 11.50 Linear scale, or 'inductosyn'

Linear scales, also known as 'inductosyns', are frequently used for the precision measurement of linear position (*Figure 11.50*). They are especially prevalent as a position sensor in machine tool slides. The scale coil is in the form of a linear 'stator' and the scanning coils operate as movable 'rotors'. The pitch of the stator and rotor coils, typically 2 mm, is accurately formed and the two rotor coils are displaced relative to each other by 0.25 of the pitch. The unit operates in much the same way as a transformer, with the rotor coils representing the primary windings and the stator coil the secondary windings. An a.c. voltage is supplied to each of the rotor coils with a phase shift of 90° between the two signals. The voltage generated in the stator coil is the sum of the two rotor coil voltages. The scale coil output can be decoded through measurement of the phase changes induced by movement of the rotor coils. Accuracies of 0.001 mm are possible, but it is necessary to count the number of cycles generated since the output signal is periodic. The inductosyn scale was developed from the synchroresolver which was the main shaft position transducer prior to the introduction of modern digital systems.

11.3.5 Measurement of rotational speed

Rotational speed is a commonly monitored parameter in general engineering measurement applications and there is considerable variety in the types of transducer used for the purpose. The absolute digital shaft encoder is suitable, but the incremental shaft encoder is more easily adapted for the measurement of rotational speed. The output waveform from any one of the photosensitive transistors on the incremental encoder is either a square pulse train, or a sinusoidal like signal. Measurement of the output signal frequency can then be simply converted to the rotational speed. The signal frequency can be determined in software using the timing facilities available on microcomputers. Alternatively, the signal can be processed through a frequency-to-voltage converter, to produce a d.c. output voltage which can be calibrated in direct proportion with the rotational speed. The latter method represents a more complete speed-measuring sensor which facilitates easier interfacing to a computer. Frequency-to-voltage converters are readily available in a range of integrated circuit forms.

The frequency-to-voltage converter also forms an integral part of many other types of speed-measuring transducer. One fairly common speed-measuring device is the magnetic pick-up. This passive transducer, requiring no external power supply, responds to the movement of ferrous parts past its pole piece. *Figure 11.51* shows the basic configuration which requires the transducer to be in close proximity to a toothed wheel. The output signal is approximately sinusoidal and a

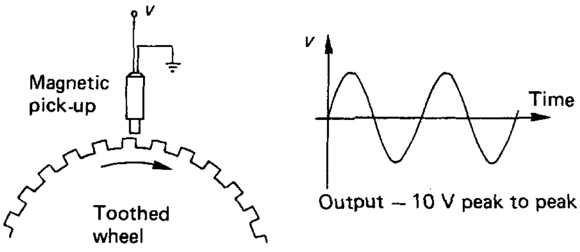


Figure 11.51 Magnetic pick-up speed sensor

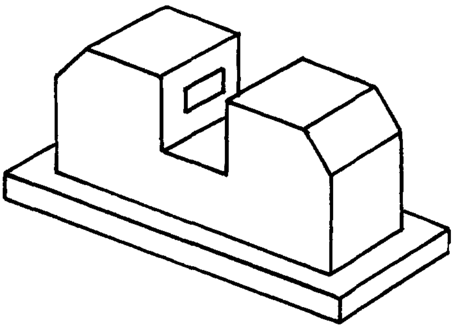


Figure 11.52 Slotted opto-switch

frequency-to-voltage converter completes the signal processing for adaptation as a speed sensor.

Optical sensing devices are also fairly common and these include slotted opto-switches (*Figure 11.52*). These usually incorporate a LED as a light source and a photosensitive transistor as a receiving device. A toothed wheel passing between the receiver and transmitter will generate the necessary pulsed output signal. Frequency-to-voltage conversion again completes the system for speed measurement.

Optical devices containing both the receiver and the transmitter on the one unit are also available (*Figure 11.53*). These units respond to the passage of a reflective surface as shown. This particular type of optical sensor is often used in hand-held tachometers.

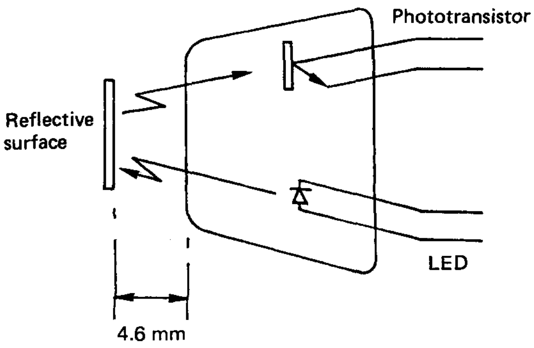


Figure 11.53 Opto-electrical sensor

Last, but not least, in the list is the well established and fairly traditional tachogenerator. These are often found in older systems where they were used as the standard transducer for speed measurement. The tachogenerator is a small permanent magnet alternator with an output voltage almost exactly proportional to speed. The voltage generated at full speed is typically of the order of 120 V. This voltage is much too high to be interfaced to a computer-based system and the output must be rectified and reduced through a voltage divider to be compatible with a digital system.

The frequency of the tachogenerator output is also proportional to speed, and alternative signal conditioning circuitry can be added to take advantage of the frequency characteristics. Conditioning circuits based on the frequency of the tachogenerator output either use a pulse counting system, or a frequency-to-voltage converter.

Due to the high cost and the high driving power required, tachogenerators are now only used on fairly large rotating machine elements.

11.3.6 Transducers based on strain measurement

In transducer technology, 'strain bridge' circuits are featured predominantly. The underlying reason for this is simply that the strain, which is essentially a displacement, can be measured as the resultant of a wide range of physical causes. The versatility of strain measurement is such that it finds application in devices adapted for the measurement of force or load, torque, pressure, flow rate, acceleration, displacement and seismic activity amongst others. The common factor involved is that the physical variable to be measured ultimately causes, by design, a deflection in an elastic member. Direct measurement of the strain associated with the deflection can then be related back to the physical variable through a calibration.

Strain-sensitive sensors are invariably incorporated into a Wheatstone bridge circuit (*Figure 11.54*). The Wheatstone

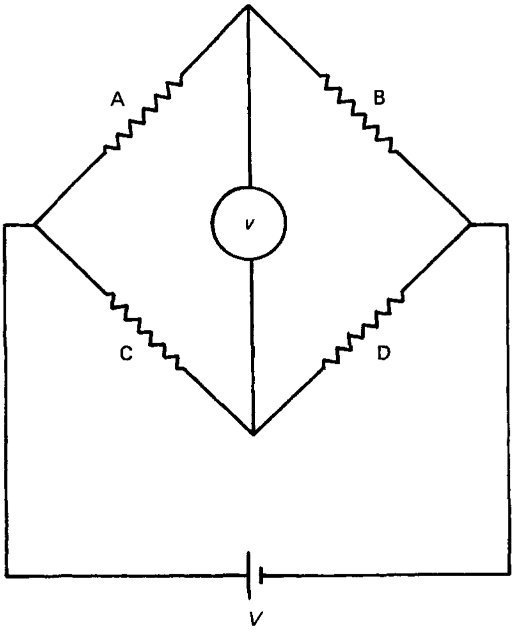


Figure 11.54 Resistive element Wheatstone bridge

bridge consists of four resistors, a voltage source and, usually, a high impedance instrument to record the bridge output voltage. When the bridge is 'balanced', no current is drawn through the voltmeter. From Ohm's law it can be shown that for balanced conditions the ratios A/C and B/D must be equal. Thus if C is a variable resistor, A may be determined if C, B and D are all known.

In most cases the sensing element constitutes one 'arm' of the bridge, while the other three arms have constant resistances, usually equal to the sensing-element resistance at particular reference conditions. These bridges normally operate in an unbalanced mode where the voltmeter reading is used to quantify the changing conditions at the sensing element, being the result of changes in the physical variable which is measured.

Assuming initially that the bridge is in balance with all elements having a resistance R, the sensing element is then subjected to a condition which causes its resistance to change to $(R + \Delta R)$. The bridge thus becomes unbalanced and a potential difference v is generated and registered on the voltmeter. Analysis of the circuit shows that provided $R \gg \Delta R$ an approximately linear relationship exists between v and ΔR:

$$v = (V\Delta R)/(4R) \qquad (11.52)$$

A detailed proof of equation (11.52) is given by Fraser and Milne.[9]

Even if the voltage recording instrument has a low internal resistance, e.g. a centre-zero galvanometer, then Thevenin's theorem shows that a near-linear relationship still applies provided that ΔR remains small.

Wheatstone bridges may also be energised with a.c., in which case they become known as 'impedance bridges'. The same principles still apply, although for equal ratios of bridge impedances only the magnitude of the voltages will be equal. The potentials may still be out of phase and a phase-sensitive instrument would indicate imbalance.

The simple resistive element Wheatstone bridge provides the basic circuit to be used with a variable-resistance sensor to form a working transducer. The sensor could be a thermistor in a temperature-sensitive system, or most commonly a strain gauge in a displacement-sensitive system.

The principle embodied in a strain gauge is that a conductor, of initially uniform dimensions, will alter its proportions when subjected to an applied stress. Since the resistance of the conductor is also dependent on the physical proportions, then the stress will also influence the resistance of the conductor.

The resistance of a conductor is given by

$$R = \beta(L/A) \qquad (11.53)$$

where β is the resistivity of the material, L is the length of the conductor, and A is the cross-sectional area of the conductor. For a cylindrical conductor, the cross-sectional area is $\pi D^2/4$, where D is the diameter of the cylinder.

Differentiating equation (11.53) gives

$$(\mathrm{d}R/R) = (\mathrm{d}\beta/\beta) + (\mathrm{d}L/L) - (\mathrm{d}A/A)$$

Now

$$(\mathrm{d}A/A) = 2(\mathrm{d}D/D)$$

Thus

$$(\mathrm{d}R/R) = (\mathrm{d}\beta/\beta) + (\mathrm{d}L/L) - 2(\mathrm{d}D/D)$$

Multiplying through by $(L/\mathrm{d}L)$ gives

$$(\mathrm{d}R/R)(L/\mathrm{d}L) = (\mathrm{d}\beta/\beta)(L/\mathrm{d}L) + 1 - 2(\mathrm{d}D/D)(L/\mathrm{d}L) \qquad (11.54)$$

where $(\mathrm{d}L/L)$ is the ratio of the change in length to the original length which, by definition, is the axial strain ε_A.

Similarly, $(\mathrm{d}D/D)$ is, by definition, the transverse strain ε_T.

Also $(\mathrm{d}D/D)(L/\mathrm{d}L) = (\varepsilon_T/\varepsilon_A) = -\nu$, where ν is Poisson's ratio for the material.

Substitution of these parameters into equation (11.54) gives

$$(\mathrm{d}R/R)(L/\mathrm{d}L) = (\mathrm{d}\beta/\beta)(1/\varepsilon_A) + 1 + 2\nu \qquad (11.55)$$

The subject of equation (11.55), $(\mathrm{d}R/R)(L/\mathrm{d}L)$, is called the 'gauge factor' and is traditionally denoted by G. The constancy of the right-hand side of equation (11.55) determines the suitability of any material to function as a strain gauge.

Since

$$G = (\mathrm{d}R/R)(L/\mathrm{d}L) = (\mathrm{d}R/R)(1/\varepsilon_A)$$

Then

$$\mathrm{d}R = GR\varepsilon_A \qquad (11.56)$$

Values of G and R are specified by the manufacturer of the strain gauge, and equation (11.56) shows that the change in resistance $\mathrm{d}R$ is directly proportional to the axial strain ε_A.

Evidently a high value of G would be advantageous in producing a high change in resistance for any input axial strain. Metallic strain-gauge materials have gauge factors of about 2, although some semiconductor materials have gauge factors in excess of 100. The commonest form of strain gauge is the bonded-foil type (*Figure 11.55*).

Bonded foil strain gauges are rigidly fixed to the strained member and are electrically insulated from it. In typical applications, the change in resistance is very small (of the order of 0.25 mΩ per microstrain) and, as a result, the measurement is very sensitive to temperature variations. Unless special temperature-compensating gauges are available, it becomes necessary to provide some other means of compensation for temperature drift. The usual method is to incorporate a 'dummy' gauge in the bridge circuit (*Figure 11.56*).

The dummy gauge is at all times subject to the same ambient temperature as the active gauge, but it is attached to a separate, unstrained piece of the same material. In this manner resistance changes due to temperature variation are cancelled out and the output voltage is a result only of applied strain on the active gauge. The extra 'compensation lead' shown in the diagram is used to negate the effect of temperature variations in the wiring to the active gauge. This is not an

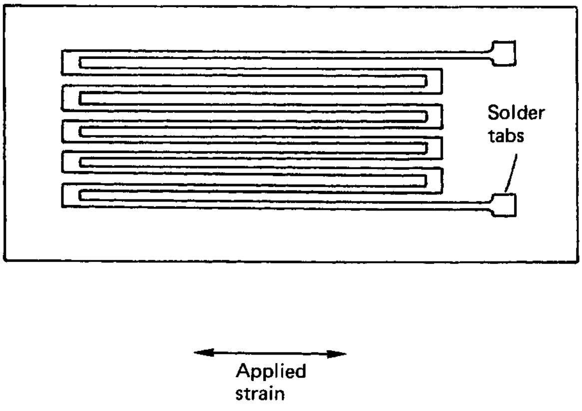


Figure 11.55 Bonded-foil-type strain gauge

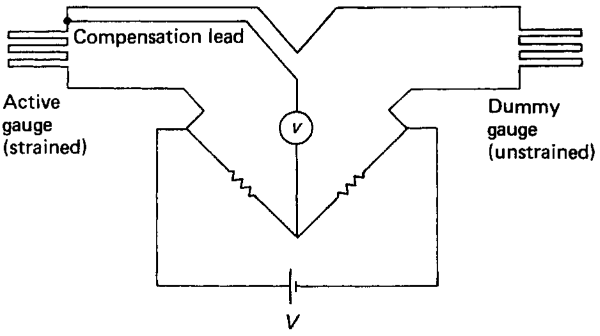


Figure 11.56 Strain bridge with temperature compensation

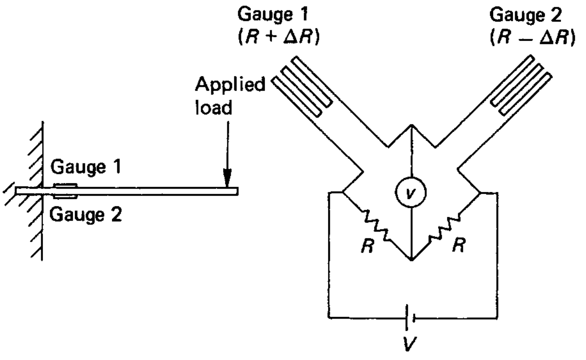


Figure 11.57 Cantilever-type load transducer

essential part of the circuit, but is advisable if the connecting cables to the active gauge are particularly long.

The sensitivity of the circuit can be increased if it is known that the flexural member is subject to equal and opposite strains on its opposite surfaces. *Figure 11.57* illustrates an application to a load transducer in the form of a simple cantilever.

The bridge output with two active gauges is:

$$v = (V/2)/(\Delta R/R) \qquad (11.57)$$

By using active gauges on all four arms of the bridge, the sensitivity can again be doubled. In either case, temperature compensation is automatic. The three forms of strain bridge described, using one, two and four active gauges are known as quarter-, half- and full-bridge circuits, respectively.

Even with four active elements in the bridge circuit the output is still measured in terms of millivolts and amplification of the output signal is essential. Furthermore, since it is almost impossible to ensure equality of the resistances within the tolerance demanded by the sensitivity of the circuit, a zero-adjustment potentiometer is also virtually essential. Other problems are associated with noise, zero drift and non-linearity, but custom-built precision amplifiers are readily available to condition output signals from resistive strain bridge circuits.

Strain bridges are often used as the sensing element in force transducers and commercial load cells. Proprietary units are available in a large variety of types and sizes for the measurement of both tensile and compressive loads. Strain bridges are also attached to tubular elements to function in the capacity of a torque meter. Flexible diaphragms are also available with a strain bridge output and these are used extensively in pressure-sensing applications.

The basic cantilevered structure depicted in *Figure 11.57* can further be adapted for use as an accelerometer, or a seismic sensitive device. In these two cases it is the force due to an acceleration which is sensed through the measurement of strain. The acceleration is then determined via Newton's law relating force and acceleration. Impulsive and shock forces may also be monitored using a cantilevered element strain bridge system.

Volumetric and mass flow rate can equally well be measured. This is done using a strain bridge which is attached to a flexural member which deflects under the drag force exerted by the flowing fluid.

The examples cited are by no means exhaustive and, indeed, it may be the measurement of the strain itself which is the primary concern. This is certainly the case in experimental stress analysis.

The strain bridge circuit undoubtedly exhibits great flexibility in its applications as a commercial transducer. Furthermore, it is probably the most extensively used sensor and circuit in many custom-built measurement applications.

11.3.7 General force, pressure and acceleration transducers

In the previous section the utility of the resistive strain bridge was highlighted in its varied applications as a transducer. In all of the examples considered it is the displacement of a flexural member which is actually measured. The measured displacement is then related back to the physical variable of interest through a calibration. Since the measurement parameter is displacement, then there is no reason why any of the other methods featured in Section 11.3.4 cannot be used as an alternative to the resistive strain bridge. In practice, and with the exception of the linear-type encoders, this is actually found to be the case. *Figure 11.58* shows four common types of commercial pressure transducer.

In all types, a change in pressure causes a deflection in a diaphragm which can be sensed using the various displacement transducers. Transducer (*a*) generates a change in capacitance between the flexible diaphragm and a central electrode. The conditioning circuitry required is quite complex and includes a capacitive bridge. The high sensitivity of the capacitance transducer, however, makes it a popular choice for many applications. The common capacitive microphone operates on these principles.

Type (*b*), which is illustrated as a differential pressure transducer, relies on the induced eddy currents generated in the coils for the sensing of the diaphragm deflection. Types (*c*) and (*d*), respectively, are miniaturised versions of the linear variable differential transformer and the variable potential divider.

Many other mechanical configurations are possible and the basic sensing principles outlined above find general application in a wide range of force, pressure and acceleration transducers.

Force, pressure and acceleration transducers based on the piezoelectric effect belong in a totally separate class. A piezoelectric crystal, e.g. quartz, produces an electrostatic charge when subjected to a compressive strain. A difficulty exists, however, since the charge is generated as a consequence of the application of the mechanical load. If the charge is subsequently dissipated, then no new charge would be generated until the loading state again changes. This feature

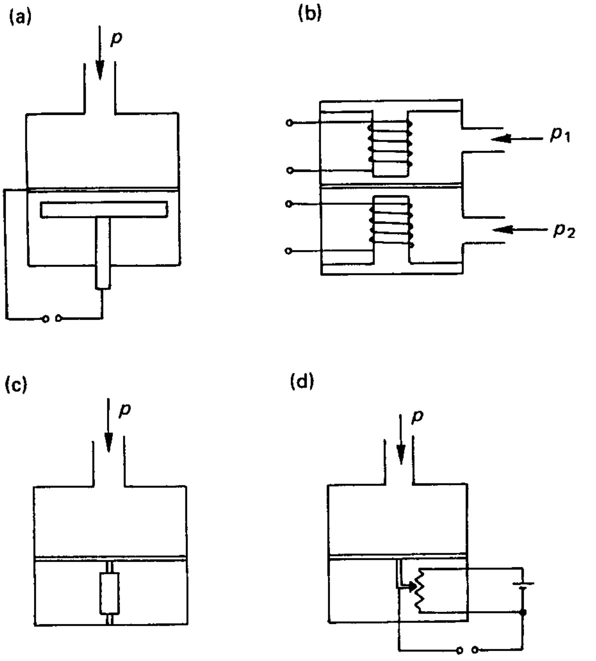


Figure 11.58 Commercial pressure transducers

makes piezoelectric devices unsuitable for static measurements. They are eminently suitable, however, for the measurement of dynamic loads. Therefore, piezoelectric transducers are used extensively to monitor the time variation of pressure within the cylinders of internal combustion engines and as an accelerometer in vibration monitoring.

Measurement of the charge, however, without dissipating it, is problematic and piezoelectric transducers require sophisticated and expensive signal conditioning.

11.3.8 Proximity sensing

Proximity sensing has many applications in manufacturing engineering ranging from simple component counting functions to the ON/OFF switching of actuators and other devices in some overall control context. In most cases the basic requirement is for a 'GO–NO GO' type measurement, and proximity switches are extensively used in these applications.

Spring-loaded microswitches are perhaps the simplest form of proximity sensor. These require mechanical contact to activate the switch and can be used effectively in the sequence control of pneumatic and hydraulic cylinders. In typical operations, the movement of a ram is initiated and allowed to continue until a microswitch is tripped by a collar attached to the ram. The microswitch then perhaps activates a solenoid operated control valve to reverse, or cut off the supply to the ram. Such systems can be extended as required to build up a fairly simple automated sequential process.

As the name might suggest, however, proximity switches are basically non-contacting devices. The magnetic pick-up and the slotted opto-switch, described for rotational speed measurement in Section 11.3.5, are both proximity devices and they can be used directly in this capacity. Many inductive and capacitive devices are also designed to function as a proximity switch. The most common type of proximity switch sensors nonetheless are the devices which operate on the 'Hall effect'. The Hall effect, named after its discoverer, is the phenomenon whereby an electron passing through a semiconductor material, and in the presence of a magnetic field, will experience a force which will deflect the electron from its normal path. The effect is similar to that used to position the light spot on a cathode ray oscilloscope, but on a much smaller scale. Modern Hall-effect sensors are supplied as complete integrated circuits with a convenient voltage output. A small permanent magnet is also usually supplied for attachment to the component to be sensed. Hall-effect sensors have the advantages of high-speed operation, no moving parts and are free from switch 'bouncing' effects.

All the above devices essentially rely on close proximity (a few millimetres or so) for their actuation. For longer range proximity sensing, light-beam transmitting and receiving devices can be used. A typical application is in the batch counting of components moving along a conveyor belt. In this type of application, a light beam is transmitted across a space, (which may be up to about 15 m) and is continuously monitored by a receiving device. The passage of a non-transparent object interrupts the beam and triggers a voltage pulse which can be measured and recorded. Reflective type proximity sensors, with transmitter and receiver housed in the one unit, are also available.

11.3.9 Measurement of flow rate

The measurement of flow rate in process control covers a range of different transducers which can be subdivided as:

(1) pressure differential devices,
(2) rotary devices,
(3) vortex shedding devices, and
(4) non-intrusive devices.

Pressure differential devices include Venturi-meters, orifice plates, flow nozzles, or any similar device which presents a constriction in the flow path. For incompressible flows, the constriction causes the fluid to accelerate with an attendant drop in pressure. The operating principle of these devices is based on two fundamental laws of fluid dynamics. These are the equation of mass continuity and the steady flow energy equation, or Bernoulli equation. Application of these laws across any type of constriction results in an equation of the form:

$$\text{Flow rate} = B\sqrt{\Delta p} \qquad (11.58)$$

where B is, in general, an experimentally determined constant, and Δp is the measured pressure differential across the constriction.

The measurement of Δp is performed with a suitable differential pressure transducer and, provided B is a well behaved 'constant', the flow rate can be accurately calibrated against the square root of the output voltage from the pressure transducer.

If the fluid is a gas it is usually necessary to measure the local gas temperature and the absolute barometric pressure in addition to the meter differential pressure. The common differential pressure flow metering devices are shown in *Figure 11.59*.

Rotary flow meters may either be of the positive displacement type or the turbine type. In either case it is the speed of the rotational element which is measured and converted into an equivalent flow rate. In positive displacement meters, the total number of revolutions is also counted and used as a measure of the cumulative volume of flow. *Figure 11.60* illustrates three contemporary rotary-type flow meters.

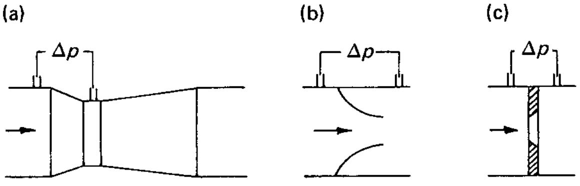


Figure 11.59 Differential pressure flow meters: (a) Venturi; (b) flow nozzle; (c) orifice

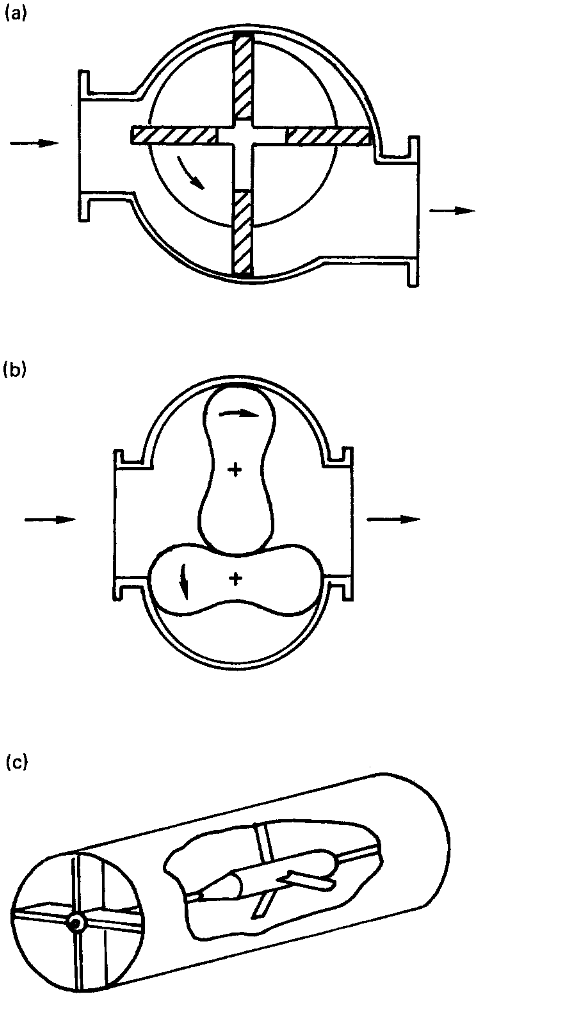


Figure 11.60 Rotary type volumetric flow meters: (a) rotary vane; (b) lobed rotor; (c) turbine

The rotary-vane and lobed rotor types are positive displacement meters in that, for each revolution of the rotor, a fixed quantity of fluid is passed through from inlet to outlet. The turbine type meter is not a positive displacement device.

In most rotary-type flowmeters, the rotational speed is detected with some form of proximity sensor which responds to the passage of a vane, lobe or blade as the case may be. A frequency-to-voltage converter is normally used in the signal conditioning process to produce a steady output voltage in proportion to the volumetric flow rate.

Vortex shedding flowmeters are comparative newcomers in the field of flow measurement, but they have become well established as a general purpose flow monitoring device. The operating principle stems from the fact that when a 'bluff' body is exposed to an oncoming flow, the fluid particles cannot follow the severe surface contours of the body. The flow therefore breaks away from the body and periodic vortices, or eddies, are shed alternately from the upper and lower extremities of the body (*Figure 11.61*).

Over a reasonable range, the periodicity of the vortex shedding is in direct proportion to the fluid velocity. A measurement of the shedding frequency can, therefore, be related to the mean velocity and hence also to the flow rate. The vortex shedding frequency is sensed by means of a variety of methods. These include pressure variation with a piezoelectric pressure transducer, temperature variation with a heated thermistor and the interference of an ultrasonic beam transmitted across the wake behind the body. Vortex shedding flowmeters have no moving parts and can be used with any fluid. The output signal is quite noisy, however, and the linear operating range might be too restrictive in some cases.

The last group of flow metering devices are the non-instrusive types. These are by far the most expensive flowmeters, but they have the advantage of leaving the flow undisturbed with no attendant loss in pressure.

Laser–Doppler anemometry is a non-intrusive method which transmits a low power laser beam into the flow. The light scattered by minute particles flowing with the fluid transmits a signal back to a receiver and a measurement of the Doppler shift frequency can be used to determine the local fluid velocity. The Dopper shift frequency is also similarly used in some of the proprietary types of ultrasonic flow meters.

Many other ultrasonic flowmeters operate on a beat-frequency measurement. These meters incorporate two separate ultrasonic transmitter/receiver probes which pass signals along opposite directions across the flow path (*Figure 11.62*).

Due to the flow of fluid, the transit times of the two signals are unequal and, by multiplying the signals together, a beat frequency can be measured. The beat frequency is proportionally related to the fluid velocity, and its measurement can therefore be calibrated for volumetric flow rate.

The electromagnetic type of non-intrusive flowmeter is worthy of mention due to its wide usage in industrial applications. These flowmeters require the fluid to be conducting and their operating principle is based on Faraday's laws of electromagnetic induction. A magnetic flux is generated in a direc-

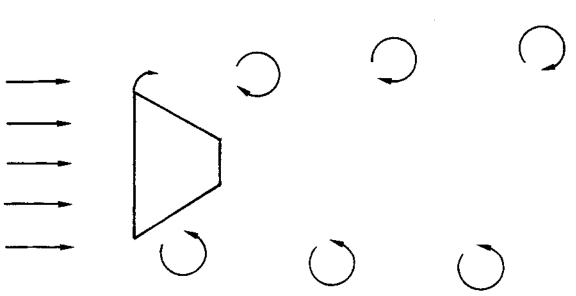


Figure 11.61 Vortex shedding from a bluff body

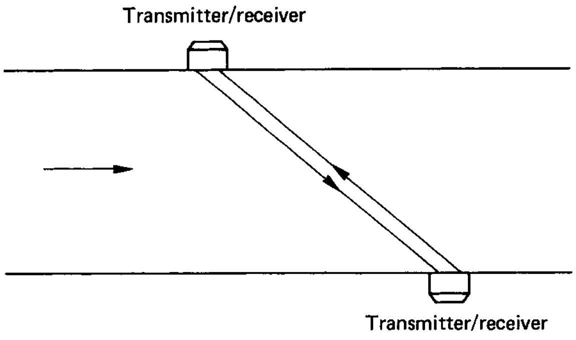


Figure 11.62 Ultrasonic 'beat frequency' flowmeter

tion perpendicular to the flow. As the flow constitutes a moving conductor, then a measurable voltage is generated between two electrodes positioned perpendicularly both to the magnetic field and the flow path. The magnetic field is energised with an alternating supply to prevent electrolysis and other problems associated with magneto-hydrodynamic action.

11.3.10 Temperature measurement

The measurement of temperature is often required in system monitoring and general process-control applications. For digital-based systems, thermoelectric devices are commonly utilised to generate the output signal in voltage form. The thermoelectric effect is a consequence of the fact that the number of free electrons which can exist in a metal is a function both of the temperature and the composition of the metal. For practical purposes it is sufficient to know that, when two dissimilar metals are joined together at their ends, a potential difference will exist if the two ends are held at different temperatures. This physical phenomenon provides the basis of a voltage-output temperature transducer (*Figure 11.63*).

In wiring up to the voltmeter a third metal (usually copper) has been introduced into the circuit. This has no adverse effect provided that the temperatures at the two copper connections are both equal.

The thermocouple, however, is a differential rather than an absolute temperature measuring device. To function as a temperature sensor, therefore, the temperature at one of the junctions (the reference junction) must be known. The most convenient reproducible reference temperature available is that of melting ice at, or near, standard atmospheric pressure. This is 0°C, and a range of specially selected materials has been rigorously calibrated against this reference to be used as standard thermocouple combinations in industrial applications. A popular pair is nickel–chromium (Chromel) and nickel–aluminium (Alumel). These two metals used in combination give the standardised '*K*' calibration, approximately 4 mV per 100°C temperature difference.

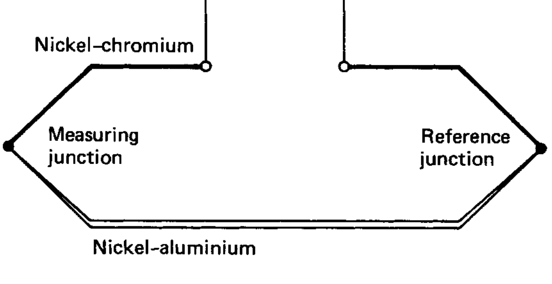


Figure 11.63 Basic thermocouple temperature transducer

The reference temperature need not be 0°C but can be any other known temperature, often room temperature (15–20°C). In these cases the reduced potential due to the higher reference temperature can be compensated for by adding the potential which would exist between 0°C and room temperature. This introduces a small error, however, since the calibrated relationship between temperature and voltage, although very near, is not quite linear. The output voltage, in addition, must be amplified to be compatible with a digital interface. Purpose-built amplifiers are readily available, however, particularly for K-type thermocouples, and these also usually incorporate a built-in ice-point compensation circuit.

If high accuracy and long-term stability are required, then a 'platinum-resistance thermometer' may be used as an alternative to the thermocouple. The temperature–resistance characteristics of the platinum-resistance thermometer are standardised in BS 1904: 1981.[10] In practice, the platinum sensor coil ultimately forms one arm of a Wheatstone bridge and matched compensating leads must be used in wiring up to the bridge (*Figure 11.64*).

If absolute accuracy is not so important, then thermally sensitive resistors, or 'thermistors' as they are generally called, can be used. Thermistors are metal oxide semiconductors with a large, usually negative, temperature coefficient of resistance. Thermistors are available in bead, rod, disc, washer and probe form. They are extremely sensitive and have a very fast transient response. However, the resistance–temperature characteristic is very non-linear and this has been a significant limitation on their general use in industry. The non-linearity is much less of a problem in a microprocessor-based system, since the non-linear effects can be easily accommodated in software, and thermistors are being increasingly used over much larger ranges than before. Thermistors with positive temperature coefficients of resistance are also available and these have particular application in over-temperature protection circuits.

All the temperature sensing devices so far considered have an upper limiting temperature of about 1000°C. The S-type thermocouple can operate up to 1700°C, but for higher temperatures, radiation-detection devices must be used. Radiation sensors incorporate a pyroelectric element which functions as a radiation-sensitive capacitance. The voltage developed across the element decreases as the frequency of the incident radiation increases. Commercial temperature-sensing devices are used in the steelmaking industries and in intruder alarm systems. The basic sensor principle can also be adapted to function as a gas detector, and it is used as such to measure the presence and density of carbon monoxide, carbon dioxide and methane in gas flows.

11.3.11 Level sensing

The simplest form of level sensor is the float-type switch. These switches are either ON or OFF depending on whether the liquid level is above or below the float. A series arrange-

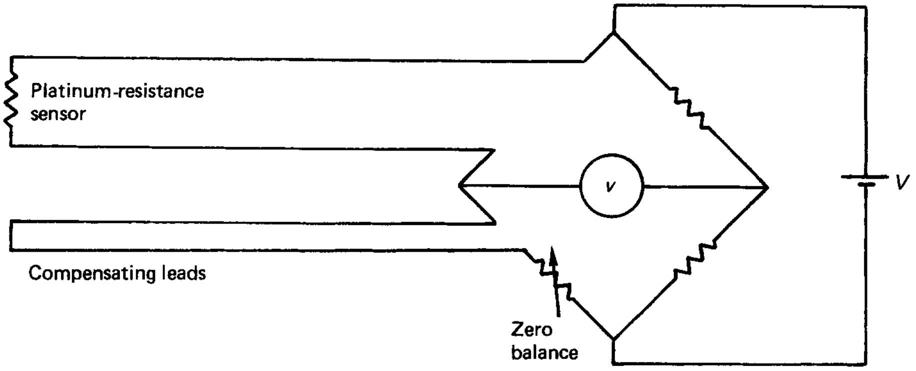


Figure 11.64 Platinum-resistance thermometer

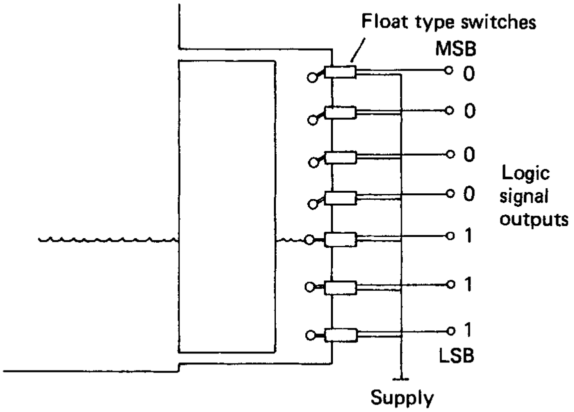


Figure 11.65 Digital output level transducer. MSB, most significant bit; LSB, least significant bit

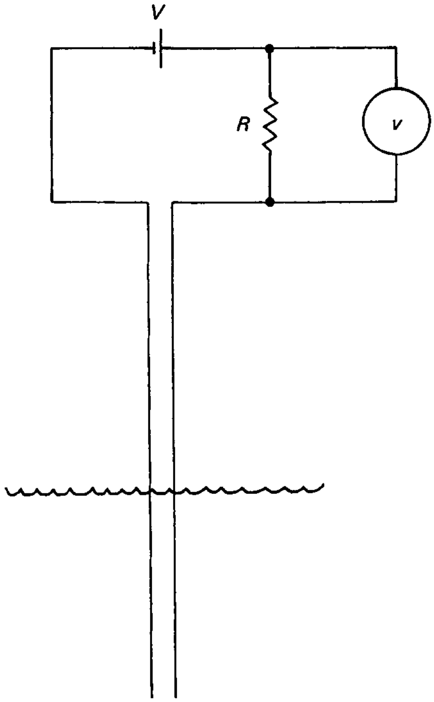


Figure 11.66 Variable output level sensor

ment of float switches (see *Figure 11.65*) could be adapted to operate as a digital level transducer.

Other digital output level transducers, similar in principle to that shown in *Figure 11.65*, use capacitive, inductive, or ultrasonic sensors to detect the liquid level. Ultrasonic and optical sensors are also used in level devices which measure the transit time of a beam reflected off the liquid free surface.

A variable output, but non-linear, level sensor is shown in *Figure 11.66*. Provided the liquid is conducting, the resistance between the two unconnected wires will depend on their depth of immersion, i.e. the liquid level. The voltage measured across the fixed resistor is inversely proportional to the depth of immersion of the two probe wires. Commercial sensors of this type, marketed as wave monitors, use an a.c. supply to prevent the adverse effects of electrolysis.

11.3.12 Bar-code readers

A form of optical data input which is finding increasing application is the bar-code reader. The code consists of a series of black and white vertical lines which are printed onto the object (*Figure 11.67*).

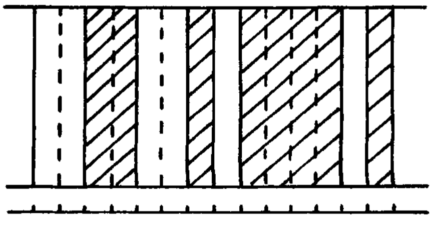


Figure 11.67 Section of a bar code

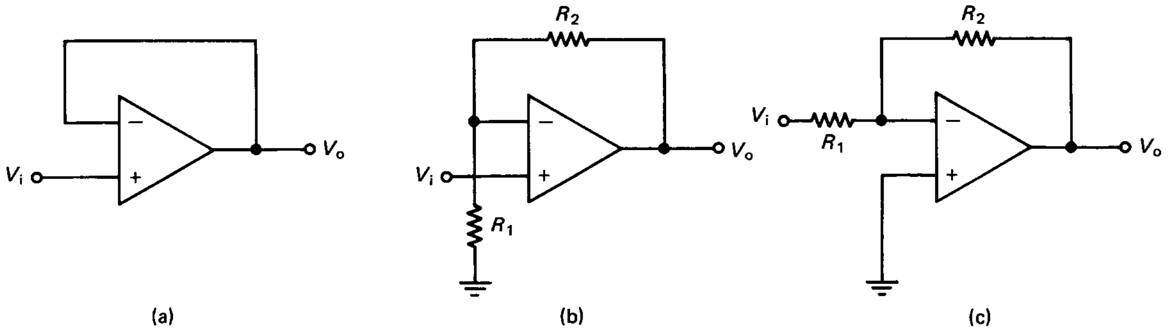


Figure 11.68 Operational amplifier circuits: (a) non-inverting unity gain; (b) non-inverting gain $V_0 = [(R_1 + R_2)/R_1]V_i$; (c) inverting gain $V_0 = -(R_2/R_1)V_i$

The code is read by an optical sensor which incorporates a lamp, a phototransistor and a number of optical focusing lenses. The decoding software is necessarily complex, however, since the speed at which the code is read can vary.

A typical bar code might consist of a start pattern (101), five 7-bit characters, a check sequence, a second group of five characters and an end pattern. Two consecutive black bands represent a bit value of 1, while two consecutive white bands represent a bit value of 0. The code is designed such that every character starts with a white band and ends with a black band. This ensures that every character starts and ends with a 1–0 transition. In addition, every character code includes at least one 1–0 transition within the code.

The decoding program must include a timing loop to determine the speed of reading. The timing is usually based on a count of the 1–0 transitions. With the reading speed established, the code can then easily be translated. Bar-code readers are very evident in large supermarkets and libraries, but they have applications in manufacturing stock control and automatic assembly lines for component counting and identification purposes.

11.3.13 Signal conditioning

Although the output signal from many sensors and transducers is already in the necessary voltage form, it is only very rarely that the voltage is high enough to be directly interfaced to a digital system. In most instances the transducer output will require some form of amplification prior to conversion from analogue to digital form.

The term 'amplifier' is a shortened form for the full description 'voltage amplifier'. Most amplifiers amplify voltage, but other types are encountered and these are normally given their full description, e.g. current amplifier, charge amplifier and power amplifier. Amplifiers are further classified as a.c. or d.c. The d.c. amplifier will accept a.c. inputs but the a.c. amplifier will block any d.c. input. The input signal may be either single ended with one active signal line, or differential with two active signal lines. Lastly, the amplifier may be 'inverting', with a reversal of sign at the output, or 'non-inverting'.

The commonest amplifier configuration embodies a differential input, single-ended output voltage amplifier with either inverted or non-inverted output. These amplifiers are referred to as 'operational amplifiers'. *Figure 11.68* illustrates some typical idealised operational amplifier circuits.

The important characteristics of an operational amplifier are its gain and its bandwidth. The gain, which can be as high as 10^6, is related to the combination of passive resistors which make up the external circuit. The gain is constant only over a resistricted range of input signal frequencies. Outside this range the gain is attenuated as shown in *Figure 11.69*.

The 3 dB 'cut-off' frequencies are used to define the bandwidth of the amplifier. The product of gain and bandwidth (GBW) is quoted by some manufacturers as a quality parameter. Other descriptive parameters associated with amplifiers are:

(1) *common mode rejection ratio*—the ability of an amplifier to reject differential input gain variation;

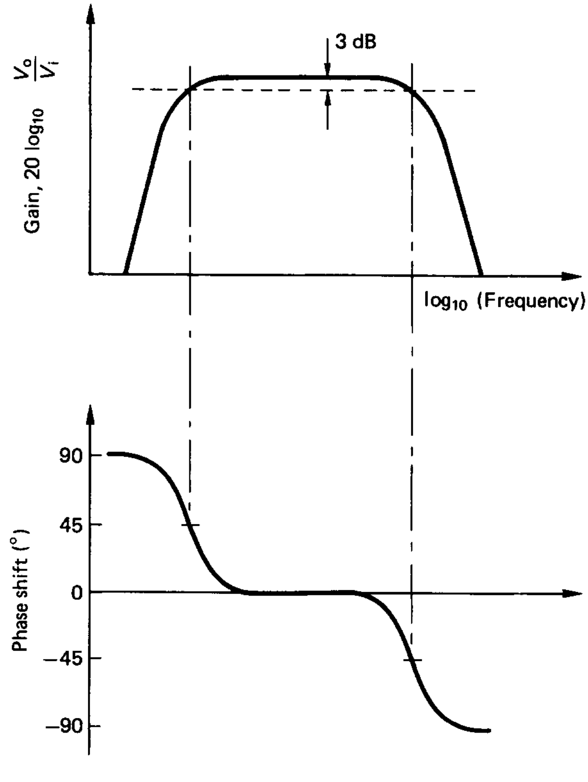


Figure 11.69 Gain and phase shift characteristics of an operational amplifier

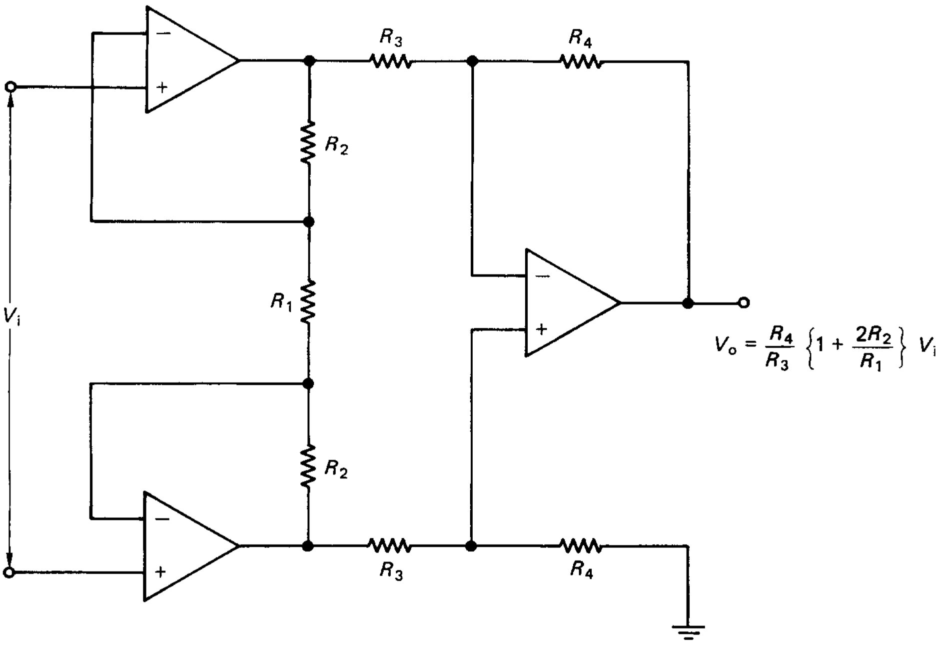


Figure 11.70 Instrumentation amplifier

(2) *offset voltage*—the voltage output attributed to input and output voltages generated by component variations;
(3) *offset drift*—temperature dependent voltage output;
(4) *non-linearity*—departures from linear input–output characteristics; and
(5) *distortion*—frequency dependent non-linearities.

Popular and inexpensive general purpose operational amplifiers are those in the so-called '741' family. For high accuracy measurements, where low drift and low noise are also essential, an instrumentation amplifier with a high input impedance and high common mode rejection ratio would normally be required (*Figure 11.70*).

The problem of impedance mismatch is frequently encountered in instrumentation systems. This occurs when a transducer having an internal resistance and functioning as a small voltage source, is coupled to a recording instrument which also has an internal resistance. If the internal resistance of the transducer is many times smaller than that of the recording instrument then the problem is insignificant. If, on the other hand, the internal resistances are of similar orders of magnitude then the error in the measurement and the loss of the signal can be quite large. The standard solution to the impedance mismatch problem is to interface an amplifier with a high input impedance and a low output impedance between the transducer and the recording equipment. Such amplifiers are usually configured to have unity gain since their primary function is the prevention of serious signal loss.

For many measurement applications, custom built special purpose amplifiers are already available in integrated circuit form. These greatly ease and facilitate the signal conditioning circuitry. Fraser and Milne[9] give a practical guide to many of the signal conditioning interfaces required for the amplification of signals from strain bridges, thermocouples and other low level voltage output sensors.

11.3.14 Analogue and digital filtering

Noise is inherently present in all physical systems where measurements are made. In sampled data systems the effect of noise, illustrated in *Figure 11.71*, can give rise to further misinterpretation in the form of 'aliases' (see also Section 11.6.4).

Discrete sampling (see above) results in an output signal which suggests that the measured variable is increasing linearly, but with a superimposed sinusoidal fluctuation. The apparent sinusoidal variation is entirely the effect of background noise and it is obviously good practice, therefore, to try and eliminate noise in the measurement system. It is perhaps fortuitous in mechanical systems that background noise is generally manifested at much higher frequencies than the frequency associated with the primary variable of interest.

The sources of noise are varied and may originate from thermoelectric effects, electrochemical action, electrostatic and electromagnetic pick-up, self-generated component noise, offset voltages and common earth loops. If the frequency content of the signal to be measured is known beforehand, then positive steps can be taken to eliminate most of the unwanted effects of noise by the inclusion of suitable filters. Filters exist in three broad categories: low pass, high pass and band pass. The gain characteristics for each type are shown in *Figure 11.72*.

A low-pass filter is one which allows the transmission of signals below a particular cut-off frequency. Signals at frequencies above the selected cut-off are progressively attenuated. The high-pass filter, in contrast, transmits only that part of the signal which is at frequencies above the cut-off value. The band-pass filter transmits, without attenuation, the signal contained within an upper and a lower cut-off value. The cut-off frequency is defined as that at which the signal attenuation is −3dB.

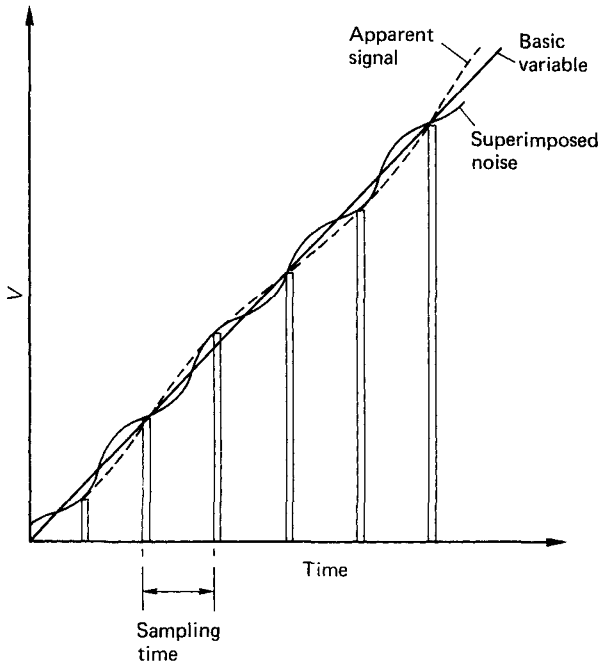


Figure 11.71 Noise generated 'aliases' in a sampled data signal

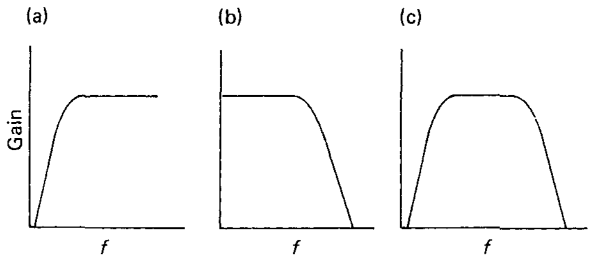


Figure 11.72 Filter performance curves: (a) high-pass filter; (b) low-pass filter; (c) band-pass filter

The simplest forms of analogue filter are those which incorporate only passive resistive, capacitive or inductive elements.

Low-pass filter The transfer function for a low-pass filter (*Figure 11.73*) is given as equation (11.14), with the gain $k = 1$. The time constant τ is equal to the product of the values of resistance and capacitance (RC). The -3 dB cut-off frequency is given by:

$$f = 1/(2\pi RC) \tag{11.59}$$

A suitable choice of resistor and capacitor, therefore, allows that desired cut-off frequency to be imposed in the signal conditioning train.

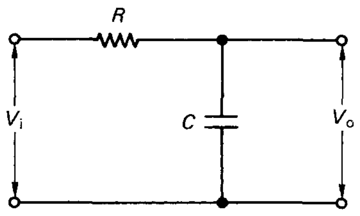


Figure 11.73 Low-pass filter

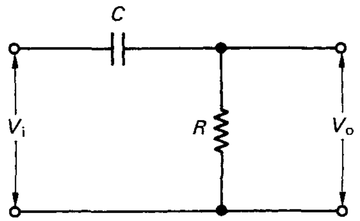


Figure 11.74 High-pass filter

High-pass filter The transfer function, in terms of the Laplacian variable, for the high-pass filter (*Figure 11.74*) is:

$$(V_o/V_i) = (sRC)/(1 + sRC) \tag{11.60}$$

The cut-off frequency is similarly selected through a judicious choice of resistor and capacitor.

Band-pass filters might be thought of as a series arrangement of a low-pass and a high-pass filter. With the appropriate combinations of resistors and capacitors, the low- and high-frequency noise components in the signal can be suitably attenuated as required.

Filters which include an amplifier in the circuit are referred to as 'active filters' and the relationship between input and output is a much more complex function of time. Williams[11] gives a comprehensive discourse on filter design.

Figures 11.73 and *11.74* illustrate the simplest forms of passive analogue filter which are used to suppress background noise. The governing equations ((11.14) and (11.60)) may equally well be expressed in terms of finite differences. In finite-difference form, the equations can be used to action the filtering process on a discrete version of the input signal. This is the basis of a digital filter which can be implemented in software and requires no external hardwired components.

The setting of the cut-off frequencies in the digital filter are achieved through adjustment of the constants appearing in the finite difference approximating function. These numerical constants are simply related to the physical time constant in the equivalent analogue filter and also the digital sampling rate.

The advantages that the digital filter has over its analogue counterpart are the ease with which the cut-off frequencies can be adjusted. The -3 dB cut-off frequency can also be set exactly, since no hardwired components, which have physical tolerance bands, are used. The digital signal may also be filtered any number of times simply by processing the data repetitively through the filtering algorithm. The disadvantage incurred is that digital filtering takes longer to perform in real

time. The end result of digital filtering on the signal is exactly the same as would be obtained using an analogue filter, i.e. any time varying signal of a frequency outside the cut-off value is subject to attenuation with a corresponding phase shift.

11.4 Microprocessor technology

11.4.1 Summary of number systems

In the manipulation of data within a computer, a two-state numbering system is used. This is termed the 'binary system' and is based on a simple ON/OFF principle. For the semiconductor integrated circuits which make up the computer system, 5 V denotes ON, or logic level '1', while 0 V denotes OFF, or logic level '0'. In practice, a tolerance band is adopted with 2.4–5 V representing logic '1' and 0–0.8 V representing logic '0'.

The microelectronic devices in the system handle the transfer of information in 1s and 0s which are referred to as 'bits', being a short form for 'binary digit'. A group of eight bits is termed a 'byte' and a number of computer systems are based on 8-bit technology with the handling of data codes as 8-bit 'words'. 16-bit and 32-bit machines are also available.

The computer operates with three numbering systems: decimal, binary and hexadecimal (often simply called 'hex'). Numerical data would normally be entered by a human operator in decimal form since this is the most familiar number system. The computer, however, must ultimately convert the decimal number into a binary code since this is the eventual form in which the number will be processed and stored. The hexadecimal system is an in-between state and represents a particularly compact method of handling binary numbers as groups of four bits.

In binary representation the only possible logic levels are 0 and 1. The base is chosen as 2, and integer numbers can be represented using 8-bit codes as shown below.

Bit number	7	6	5	4	3	2	1	0
	2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0
	128	64	32	16	8	4	2	1
	Most significant bit (MSB)							Least significant bit (LSB)

The conversion from binary to decimal is as follows:

Binary number	1 0 1 1 1 0 0 1
Giving	128 + 0 + 32 + 16 + 8 + 0 + 0 + 1 = 185 decimal

Conversion from decimal to binary is the reverse process to the above.

It is apparent that the highest number which can be accommodated in 8-bit binary notation is 1111 1111, which is equivalent to 255 decimal.

Generally, therefore, computer systems handle integer numbers in four consecutive bytes, i.e. as 32 bits. The most significant bit is used to denote the sign of the number and the resulting range of integer numbers is:

-2^{31} to $2^{31} - 1$
or
−2 147 483 648 to 2 147 483 647

Real numbers are handled in five bytes with the most significant byte representing an exponent and a sign bit and the other four bytes representing the mantissa and a sign bit. The resultant range of real numbers is 0.5×2^{127} to 2^{127}, with either a positive, or a negative sign for the mantissa.

The handling of numbers in binary notation is extremely cumbersome for a human and a short-hand notation is adopted for convenience. This is the hex system in which the binary number is arranged in groups of four bits. Four bits, which is half of a byte, is called a 'nibble'. A byte therefore consists of an upper and a lower nibble.

Since there are only 10 unique symbols in the decimal numbering system, the first six letters of the alphabet are used to denote the additional six symbols in the hexadecimal system.

Decimal	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Hexadecimal	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F

Thus, using hex notation, 8-bit binary numbers may be replaced by two hex symbols, e.g.

167 decimal ≡ 1010 0111 binary ≡ A7 hex

Higher numbers are handled similarly, e.g.

6836 decimal ≡ 0001 1010 1011 0100 binary ≡ 1AB4 hex

11.4.2 ASCII code

In the interchange of information between the constituent parts of a computer, or a peripheral device, a binary code is used to represent the alphanumeric characters. The most commonly used code for digital communication links is the American Standard Code for Information Interchange (ASCII pronounced Askey). ASCII is a 7-bit code which can accommodate 128 definable characters.

When communication takes place in a serial fashion, the ASCII code is extended to 8 bits, usually by inserting a zero in the most significant bit. In addition, one or two start bits, a parity bit and a stop bit are also included. The start bit(s) informs the receiving device that a character code follows. The parity bit provides a check that no bits have been corrupted during transmission, by ensuring that the sum of all the 1s in the ASCII group give either an even number for 'even parity', or an odd number for 'odd parity'. The stop bit, set to logic '1', terminates the transmission of the character.

The transmission rate in bits/second is termed 'baud'. Since there are 11 bits associated with the transmission of one character, a speed of 2400 baud corresponds to 2400/11 = 218 characters/second.

11.4.3 Gray code

The Gray code is just one of many binary codes in which only one of the digits changes between successive consecutive numbers. The main application is in the sensing of rotational and translational position in mechanical systems (see Section 11.3.2).

In converting from Gray to binary code, the most significant bit of the binary number B is equal to the most significant bit of the Gray code G. For all other bits, the relationship between binary and Gray is given by

$$B(n) = G(n) \oplus B(n + 1) \qquad (11.61)$$

where n denotes the bit reference number and $\oplus$ represents an exclusive-OR logic comparison.

The conversion of 1101 Gray to binary is shown in *Figure 11.75*; 1101 Gray ≡ 1001 binary.

12 Machine Tool Control Elements

G M Mair

Contents

12.1 Machine tool control system—overview 12/3
12.1.1 Open-loop control 12/3
12.1.2 Closed-loop control 12/3
12.1.3 Control-system stability 12/4
12.1.4 Controller types 12/4
12.1.5 Path control 12/4
12.1.6 Control-system hardware 12/5

12.2 Electric motors 12/5
12.2.1 D.c. permanent-magnet servomotors 12/5
12.2.2 Brushless servomotors 12/5
12.2.3 Stepping motors 12/6

12.3 The servomotor control and amplifier 12/7

12.4 Transmission elements 12/8
12.4.1 Gears 12/8
12.4.2 Screw–nut systems 12/8

12.5 Fluid power actuators 12/9

12.6 Fluid power actuator control valves 12/9
12.6.1 Control-valve specification 12/10
12.6.2 Pneumatic control valves 12/11
12.6.3 The hydraulic servovalve 12/12

12.7 Feedback transducers 12/13
12.7.1 Position and proximity sensors 12/13
12.7.2 Displacement transducers 12/16

12.8 Conclusion 12/20

12.1 Machine tool control system—overview

This chapter provides an overview of machine tool control system concepts. It allows an appreciation of the functions needed to be performed by each part of the system. It also describes these elements and explains how they relate to each other.

Machine tools may be controlled manually or automatically. Automatic machines may employ 'hard' automation in the form of cams and fixed sequence mechanical or pneumatic systems, or 'soft' automation in the form of numerical control or other easily reprogrammable systems. Although most of this chapter is concerned with the elements used in numerical control machine tools, much of what is described is common to other types of manufacturing equipment.

Cam-controlled machines are employed in mass-production systems where frequent changes in machine-element movements and their sequences are unlikely. Cams were the earliest means of automatic control and are still to be found today on automatic lathes and dedicated assembly machines. *Figure 12.1* shows some examples.

Termed 'brain wheels' on Christopher Spencer's original automatic lathe cams are still found on machine tools used for mass production. They can be used to simply trigger ON/OFF signals by activating pneumatic or electric switches. They usually control the pattern of movement of machine tool elements such as slides or cutting tools, in this case the cam profile is cut to provide the desired movement to be imparted to the element by the cam follower. The most common type of cam is the flat-plate cam cut from steel plate. This type may have rotary or linear input as shown in the figure. Other types of cam are also used, e.g. the radial closed track and the drum cams. In each case the control signal is represented by the cam surface relative to a set datum. This is translated into movement via the cam follower, one end of which rests on the cam surface and is constrained to respond to the cam movement.

Numerically controlled machine tools may be *open loop* or *closed loop*, they will also have a *path control* system specified.

An additional designation used is *servo control* or *non-servo control*. Originally, 'servocontrol' implied a system that involved amplification of a control signal, e.g. power steering in a motor car, where the manual force at the steering wheel is amplified by the servosystem before transmission to the road wheels. In machine tools this type of servo control is always present, whether the control signal is a pulse of low-pressure air to a spool valve in a pneumatic circuit, or a voltage to an amplifier in an electrical system, the signal always results in an amplified response. Thus, in the context of industrial equipment, the term 'servo control' normally has a different connotation. It usually indicates the capacity of the system or element for closed-loop control. These terms are now described in more detail.

12.1.1 Open-loop control

This type of control is found, for example, on simple machines that are under direct manual control, or cam-operated machines, or machines under computer control that are only subjected to small loads. *Figure 12.2* illustrates the principle. A signal is sent from the operator, or computer, to the amplifier and actuator. It is assumed that the machine responds in an appropriate manner to the command. In practice, the loop is often closed in a rudimentary manner, e.g. a human operator may confirm the response visually, or a microswitch might be mechanically activated when a desired position is reached. In light-duty applications, electric stepper motor and pneumatic systems operate in this way.

12.1.2 Closed-loop control

For most computer-controlled medium to heavy duty industrial equipment this is the normal method of control. It involves the use of *feedback*. The principle is illustrated in *Figure 12.3*. A control signal is sent from the controller to the amplifier and actuator, the response of the system is monitored by feedback sensors. These sensors send signals, propor-

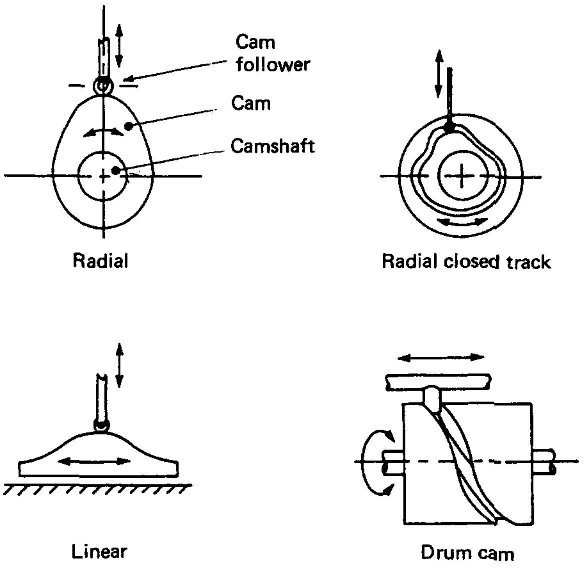


Figure 12.1 Types of cam

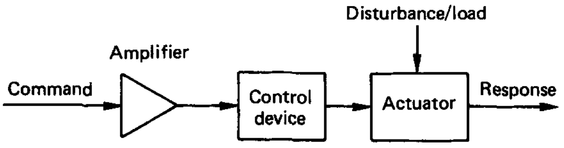


Figure 12.2 Open-loop system

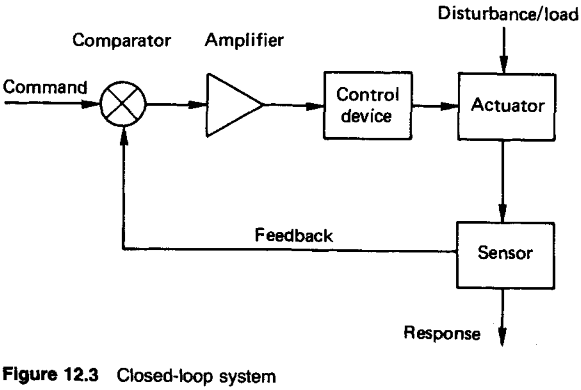


Figure 12.3 Closed-loop system

tional to the variable being measured, back to the controller. The controller compares the feedback signals with its original control signal and then modifies its output to compensate. Thus a 'loop' is formed which keeps the system under control. The British Standards Institution defines a closed-loop control system as, 'a control system possessing monitoring feedback, the deviation signal formed as a result of this feedback being used to control the action of a final control element in such a way as to tend to reduce the deviation to zero'.

12.1.3 Control-system stability

One of the factors to be considered when designing a control system is its *stability*. This is determined by the manner in which the system responds to a suddenly applied signal. As noted above, in a closed-loop system, self-correction occurs; however, if overcorrection is present then *oscillations* or 'hunting' may appear. This happens when the system overcorrects itself first in one direction and then in another. If this is too severe then the system does not settle down and in a severe case will run out of control, just like a car in a skid for which the driver overcorrects. As this condition is obviously unsatisfactory it is normal to slow the system response down by *damping*. Critical damping occurs when the system produces the desired result exactly, overdamping occurs when the system response is too sluggish, and underdamping occurs if the system oscillates a few times before settling. In an electrical system damping can be achieved by increasing the electrical resistance, this is analogous to a mechanical system, such as a shock absorber, in which oil is forced through a small orifice. Another method of preventing oscillations is the use of a large 'dead-band' region, or 'differential gap', on either side of the desired value. Thus the system stops correcting itself as soon as the feedback signal enters the dead-band region. As the dead band increases, so the settling time and precision of the system decreases.

12.1.4 Controller types

There are six common types of controller incorporating the four basic control actions of, ON/OFF, proportional, derivative, and integral control. These types are outlined below.

12.1.4.1 ON/OFF control

In this type of controller the control signal is either fully on or fully off. To prevent hunting here it is usually necessary to have a dead-band region for the error signal to pass through before switching takes place. Using a car analogy again, the engine coolant system thermostat is a good example of an ON/OFF controller.

12.1.4.2 Proportional control

This gives a smoother control action. The controller basically performs as an amplifier and provides a control signal proportional to the error.

12.1.4.3 Integral control

Here the rate of change of the control signal is proportional to the error signal. A large error signal produces a rapid increase in the control signal, a small error signal produces a slow control signal increase. A zero error produces a constant control signal. These controllers can be used when there is a constant load on the system. This is because a control signal will still be produced to counteract the load even when no error signal is present.

12.1.4.4 Proportional plus integral control

Some of the disadvantages of the previous two systems are complementary and can be overcome by combining them. For example, a proportional controller needs an error signal to be able to counteract a load, whereas the integral type can cope with zero error but often provides a slow response.

12.1.4.5 Proportional plus derivative control

In derivative control the control signal is proportional to the rate of change of the error signal. This is seldom used on its own as, without a changing error signal, the controller would produce zero output. The advantage of derivative control is that it anticipates changes in the error, thus providing a more rapid response to changes.

12.1.4.6 Proportional plus integral plus derivative (PID) control

This is the commonly used PID controller formed by a combination of the three main control actions. It provides a fast response and good stability and precision.

12.1.5 Path control

This defines how the components of the machine move. In the case of a machine tool the terminology is usually, 'point to point', 'straight line', and 'contouring'. In the case of an industrial robot, 'point to point', 'point to point with coordinated path', and 'continuous path' would probably be used. These terms are described in *Figure 12.4*.

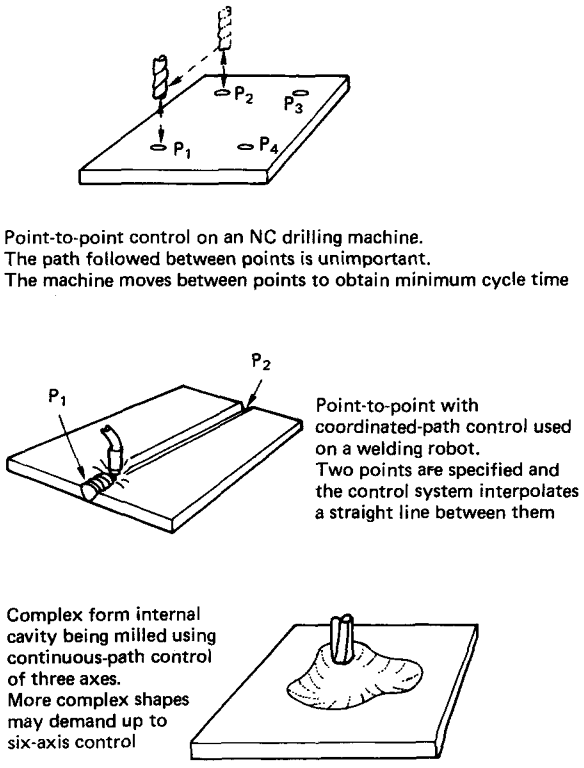


Figure 12.4 Point-to-point, straight-line and continuous-control paths

12.1.6 Control-system hardware

The control system in a machine tool is made up of a number of physical elements. Firstly there is the controller cabinet which, in a modern machine tool, will probably house the microprocessor control system or 'computer'. This contains the logic units, interfacing boards, and data-handling equipment, etc., i.e. the low-voltage electronics. In an electrically powered machine there will also be the motor control and amplifier circuits to provide the high power output to the electric motors. In fluid power machines, i.e. pneumatic and hydraulic, control of the actuators will be implemented by solenoid operated valves and hydraulic servovalves interfaced to the controller. The movement of the machine tool components is effected by electric motors of various types acting via transmission components such as gears, ballscrews, etc. In the case of industrial robots, pneumatic and hydraulic rams and rotary actuators are also used. Feedback to the controller is provided through the use of sensors such as optical shaft encoders, tachometers, etc.

12.2 Electric motors

There are essentially three types of electric motor normally used in machine tools: d.c. permanent magnet and 'brushless' servomotors for drives, and much less frequently, stepping motors for light-duty positioning. The term 'servomotor' is used to signify that the motor includes an integral feedback device, usually a tachogenerator or optical shaft encoder. A brief description of each type of motor is given below together with comments on their characteristics and advantages and disadvantages.

12.2.1 D.c. permanent-magnet servomotors

The basic construction of a d.c. permanent magnet servomotor is shown in *Figure 12.5*. Permanent magnets are used to provide a permanently energised magnetic field. These magnets are increasingly of the rare-earth type, e.g. samarium cobalt and neodymium iron boron, which provide a relatively high flux density in relation to the earlier Alnico or ferrite magnets. This reduces size and weight in relation to torque produced. The magnet segments (there may be four, six, or eight poles) are fixed to the steel outer housing by adhesive or bolts. This type of construction, as opposed to separately excited types, facilitates control, improves reliability, increases efficiency, reduces heating, and simplifies construction. Since the field flux is constant the torque generated is directly proportional to the armature current, i.e. speed changes can be effected by altering the armature voltage. There is a current regulator incorporated in the system to compensate for speed drops that may occur due to static and dynamic loads on the motor.

The cylindrical type of motor is the most common in machine-tool construction. However, where fast response and high acceleration are required, and there is little danger of overloading, then motors with a disk rotor may be used. The disk-shaped armature is lightweight as it is not composed of iron. The conductors are stamped from sheet copper and lie in layers on the disk which rotates between the exciter and magnets. It is important that these motors are not overloaded as they as susceptible to burning out due to excess current.

In situations where continual high torques and speeds are required, the d.c. permanent magnet servomotor is likely to prove suitable. It provides smooth operation over a wide range of speeds and is especially efficient at high speeds, it is also relatively inexpensive. However, the brushes in a d.c. motor create problems. For example, they restrict its use in applications that require a high frequency of starts, stops, and direction changes; the possibility of sparking at the brushes prohibits its use in flammable or explosive atmospheres; as the brushes and commutator wear small particles are created, this makes its use in 'clean' environments difficult; and of course maintenance and replacement of the brushes is regularly necessary. For these reasons, and for heavier duty applications, brushless motors are increasing in popularity.

12.2.2 Brushless servomotors

Though remaining a more expensive solution than the d.c. servomotor, the brushless servomotor has a number of advantages. Since there are no brushes, maintenance costs due to brush and commutator wear are reduced and such motors can be used in some hazardous areas. They are quiet and reliable. Rotor inertia is reduced, they are capable of very high speed

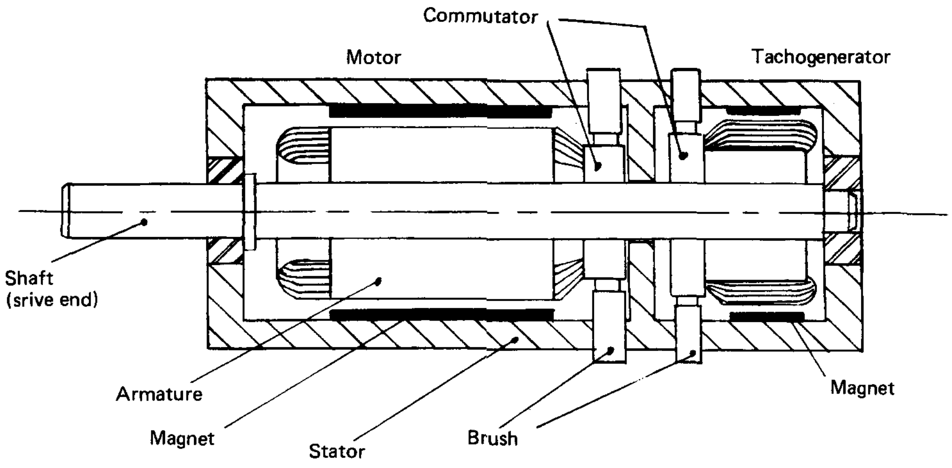


Figure 12.5 Basic construction of a d.c. permanent-magnet servomotor

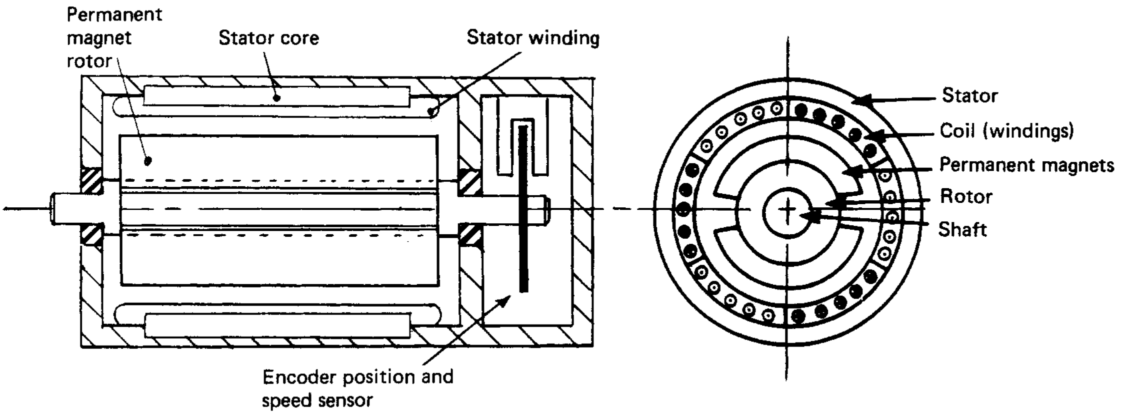


Figure 12.6 Basic construction of a 'brushless' servomotor

operation, and high peak torques are obtainable. They have smaller size and weight-to-power ratios than other types. Heat dissipation is better as heat is more easily lost from the stator than the rotor. However, despite the basic motor being of simple construction and, therefore, less expensive, the control system is relatively complex making a full system more expensive than a conventional d.c. alternative.

The basic construction of a brushless motor is shown in *Figure 12.6.* In this type of motor, permanent magnets are mounted on the rotor and the windings, or coils, are wound on the stator, i.e. the inverse of the d.c. permanent-magnet servomotor design. The motor operates in a synchronous manner, that is to say the rotor rotates in synchronism with the frequency and direction of the current supplied to the coils. The field coils in the stator are electronically commutated by switching the current from one coil to another. By controlling voltage and frequency, the motor speed under load is also controlled, i.e. the rotor speed is the same as the rotational frequency of the magnetic field. The synchronisation of the rotor speed and rotating magnetic field is essential. This is obtained by continuously monitoring the angular position of the rotor by the use of precision feedback elements such as shaft encoders or resolvers. These position data are used to control the electronic commutation. These 'brushless motors' may be a.c. or d.c. The physical construction of each type is similar, it is the form in which the current is supplied to the coils that differs. In the a.c. type a polyphase sine wave a.c. is used, in the d.c. type it is a d.c. pulse. After conditioning of the signal in both types and modification of the waveforms and pulses, there may in fact be very little difference. Thus the term 'brushless motor' is used to cover both types.

12.2.3 Stepping motors

On modern machine tools, high feed and traverse rates in conjunction with relatively high torques are required. Stepping motors are not particularly suited to these conditions. However, since there are often opportunities for using them on special-purpose and custom-made machines, they are discussed briefly here. The basic construction of a stepping motor is shown in *Figure 12.7.* The motor is composed of a multipole permanent-magnet rotor and a stator with a number of poles energised by windings. By passing current through these windings the poles on the rotor are caused to align with the stator field. The stator poles are energised sequentially so that a stepping action is produced as the rotor follows the rotating field. There are three types of stepping motor:

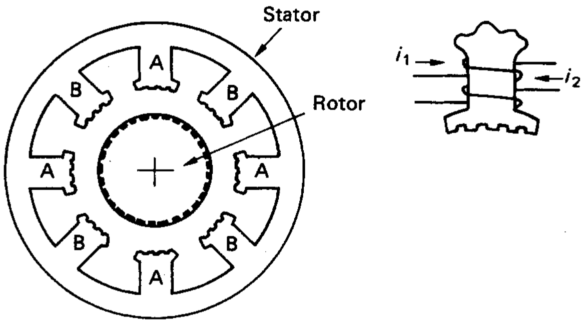


Figure 12.7 Basic construction of a stepper motor. The windings have been omitted for clarity; there is one set of windings for poles A and one set for poles B. i_1 and i_2 produce the north and south stator poles, respectively

(1) *permanent magnet*—normally with large step angles of 20° to 90°;
(2) *hybrid*—these have smaller step angles between 1.8° and 2.5° and provide higher dynamic and holding torque. Both permanent-magnet and hybrid types exhibit a detent torque when power is removed, this retains rotor position against small loads; and
(3) *variable reluctance*—this type has step angles between those of the permanent-magnet and hybrid types, lower torque and higher speed than the permanent-magnet type, but does not exhibit a detent torque.

The advantages of stepping motors are as follows. They have a simple, low-cost construction. Under small load conditions they can directly convert digital instructions into incremental rotational displacement, with no necessity for feedback. They are very reliable and negligible maintenance is necessary. Torque is good at low speeds. The disadvantages of stepper motors prevent them being widely used on heavy duty machine tools. Some disadvantages are as follows. Their resolution and stepping frequencies are too low, this results in low feed and traverse rates on the machine tool axes. Although torque is good at low speeds, maximum torque is

usually poor and this results in poor acceleration. Where medium loads are experienced then feedback sensors must be built into the system, otherwise rotational steps might be missed. Thus low-speed, high-torque conditions requiring short, fast, repetitive movements are suitable for stepping motor application.

12.3 The servomotor control amplifier

Suppliers of electric servomotors normally provide complete control systems including the control amplifiers. In the case of d.c. permanent-magnet and brushless motors these are now mostly transistor types using pulse width modulation. Amplifiers incorporating thyristors are not now very common due to poorer regulation capabilities. Stepping motors are driven via amplifiers containing purpose-built integrated-circuit chips. These units can take direct input from microprocessor-based digital controllers. There are at least three inputs to each amplifier unit: one carries the start/stop signal, one the rotational direction signal, and one the drive pulses. The output from the unit is the power pulses to the stator windings on the motor. Since, as mentioned above, stepping motors are not widely used on machine tools, and thyristor control for d.c. motors is not now common, the remainder of this subsection briefly considers the principle of the pulse-width-modulation method of control.

The control amplifier has essentially two functions, i.e. to supply power to the motor to cause it to run, and to control this power so as to maintain the required speed. The low-voltage command to the amplifier is usually in the region of −10 V to +10 V. Thus a 0 V command means standstill, −10 V means full speed reverse, and +10 V means full speed forward. If the input voltage is changed from −5 V to +5V, the motor must respond by changing from half speed reverse to half speed forward. The motor cannot do this instantaneously, it must overcome inertia and decelerate and stop before reversing direction and accelerating to the new speed. To control a motor efficiently in this way *four-quadrant control* is necessary. This allows control under the following conditions.

(1) *Motoring forward*—In this case the motor is accelerating the load and rotating in a clockwise direction, i.e. the motor is driving the load.
(2) *Braking forward*—In this case the motor is decelerating the load and rotating in a clockwise direction. The torque is reversed from (1), i.e. the load is driving the motor forward.
(3) *Motoring reverse*—In this case the motor is accelerating the load and rotating in an anti-clockwise direction. Torque is reversed from (1), i.e. the motor is driving the load in reverse.
(4) *Braking reverse*—In this case the motor is decelerating the load and rotating in an anti-clockwise direction. The torque is the same as in (1), i.e. the load is driving the motor in reverse.

To implement this control, the previously mentioned pulse-width-modulation amplifier is used. The detail of the construction of each amplifier depends on the type of power supply, type of power output, and type of motor being used. For this reason it is simply the principle of operation that is shown here.

Initially the three-phase mains supply is rectified to produce a constant d.c. voltage which will provide the high power supply to the motor. This d.c. voltage is broken up by the pulse-width-modulation unit into pulses of the desired width and frequency. The pulses occur at frequencies of several kiloherz and are filtered by the inductance of the motor coils, thus preventing torque pulsation. The principle of operation is shown in *Figure 12.8(a)*, only two pairs of transistors are shown for simplicity. The transistor controller switches the power transistor pairs on or off, depending on the input signal it receives from the system controller. Current is allowed to flow in the forward direction by switching one pair of transistors, say AA′, on and the other pair, BB′, off. Current flow is reversed by switching the state of both pairs, there is usually a very small interval during which all transistors are off, this prevents them being damaged. In *Figure 12.8(b)*, a zero net

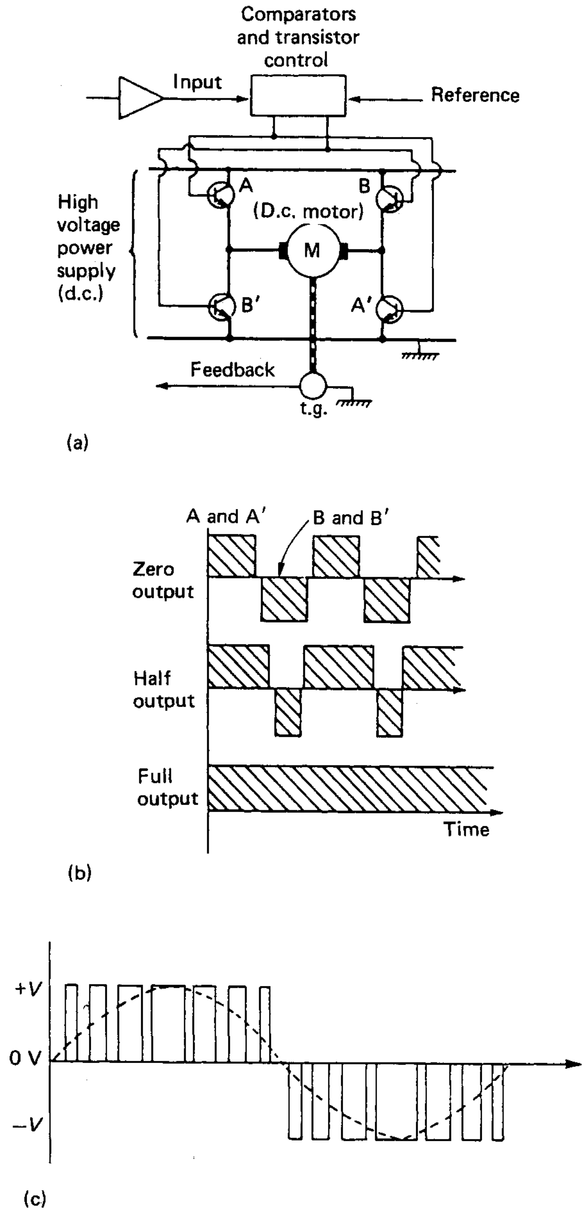


Figure 12.8 Pulse-width-modulation control

torque is produced by the motor when both pairs of transistors are switched on for an equal duration, a forward torque is produced when pair AA′ is switched on for longer than pair BB′, and when pair AA′ is fully on and BB′ fully off then the motor draws maximum current in the forward direction, but speed control is not possible. *Figure 12.8(c)* shows the pulse-width-modulation waveform for one type of brushless motor which requires pulses in a form that provides an approximation of a sine wave.

12.4 Transmission elements

In order to transmit movement and force from an actuator to the axis of a machine tool, industrial robot, or other piece of machinery, some sort of transmission system is required. These systems are mainly associated with electrically powered machines which are now the most popular in industry. In a few cases, particularly in robotics, direct electric drives are employed that require no transmission elements; however, these are still uncommon. Transmission systems also allow speed reductions or increases between the actuator and machine axis, and the transformation of rotary motion to linear, and vice versa. The most common transmission elements used are gears and screw–nut systems, these are described below. There are also special gear systems such as harmonic and cycloidal drives, and elements such as timing belts for light duty synchronous transmission.

12.4.1 Gears

Gears comprise the most ubiquitous type of transmission system; they are to be found in almost all machines. Gears provide good torque transmission, can be purchased 'off the shelf' if necessary, are reliable, and can be produced in many types to suit a variety of applications. For example, spur gears are the simplest type and have their teeth parallel to the gear shaft axis. They are relatively inexpensive but are not suited to high loading and they can be noisy. For these reasons, helical gears are popular as they have a greater area of tooth contact and the load is distributed over two or more teeth at any moment. This means that, although more expensive, they can transmit higher loads and are smoother and quieter in operation. However, the geometry of helical gears causes end thrust on their shafts, which in turn necessitates the use of thrust bearings. At some further cost this problem can be overcome by the use of herringbone or double helical gears. For non-parallel shafts, crossed helical gears can be used, although they have a lower load capacity. For shafts whose axes intersect at right angles, bevel gears are used. The straight bevel is analogous to the spur gear and the spiral bevel to the helical gear. For high reduction ratios on right-angled shafts, worm gears can be used; these also have a high load capacity. To transform rotary motion to linear, and vice versa, rack and pinion systems are used. The rack is effectively a gear of infinite diameter. The rack may be pushed by a linear actuator, so causing rotation of the pinion, or the pinion may be driven by a motor so causing either linear movement of the rack or the pinion to be driven along the rack. These gears are shown in *Figure 12.9*.

'Backlash' is the term used to describe the lost movement between the input and output shafts due to clearance between the gear teeth, it may also be referred to as 'play' or 'give'. It is usually unwanted but often some has to be built in to prevent binding, jamming, overloading, overheating, and noise such as high speed whistling. Thus backlash is evidenced by non-movement of the output shaft for an amount of time when the input shaft changes direction. As well as built-in backlash, unintentional play can develop in a system due to wear of the contacting surfaces of the components. For high-precision work, such as is found in machine tools and industrial robots, backlash generally needs to be kept to a minimum. Various techniques can be applied to achieve this, each one suited to particular conditions; a few are noted here. The simplest method is to use spring-loaded split, or 'scissor' gears. Here the gear is in two halves, i.e. it is split along a plane parallel to its face and at right angles to its axis. One half is fixed to the gear shaft, the other is spring loaded and free to slide over the face of the fixed half. Thus, when mated with another gear, the springs force the two halves to rotate relative to each other until no backlash exists between the mating gears. This method is only suitable for low-torque applications. For higher torque situations split gears can be used. The gears are locked in position, after adjustment, by using screws or rivets. Another method is to use adjustable centres where the centre distance between gear pairs is manually adjusted before clamping the gear in position. Spring-loaded centres may also be used for lighter loads. Alternatively, spur or helical gears can be cut at a slight angle to produce tapered teeth. The gears can then be moved in an axial direction relative to each other and thus reduce or remove backlash.

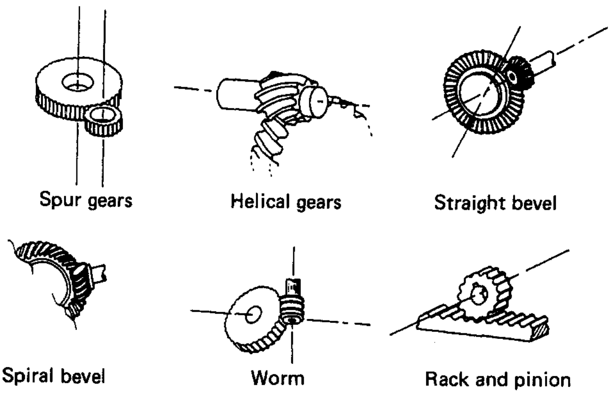


Figure 12.9 Gear types

12.4.2 Screw–nut systems

Also extremely common, screw–nut systems are found on almost all machine tools, many industrial robots, and other pieces of manufacturing machinery. While gearing is used to effect speed reduction and variance, screw–nut systems are used to translate rotational movement into linear movement of the machine elements. The screw is normally driven by an electric servomotor, either directly or through a gear system. In a machine tool the nut is fixed to one member, e.g. the machine table, and the screw to the other, e.g. the machine bed. As the screw rotates the nut is driven linearly along the axis of the screw, thus causing the table movement. This type of transmission system provides reliability, rigidity, repeatability, and resistance to vibration.

It is usual for ball or roller screws to be used because, in machine tools, it is important that smooth operation of the elements occurs. One factor that can prevent this is the phenomenon of stick–slip. This occurs at slow speeds due to the system experiencing fluctuations between static and sliding friction which in turn is caused by the lubricant film spora-

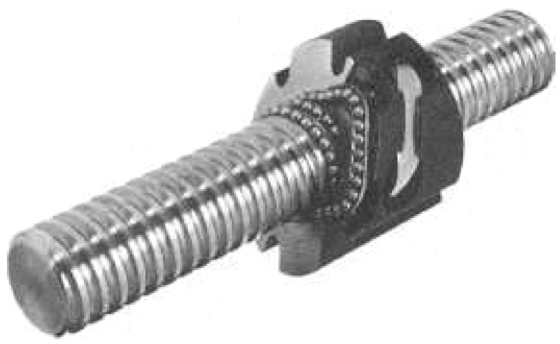


Figure 12.10 Ball nut–screw arrangement. (Courtesy of Transrol SKF Ltd)

dically breaking down. One means of overcoming this is to use rolling friction elements to reduce the coefficient of friction; this also has other beneficial effects such as reducing temperature rise at high speeds and reducing power requirements. A section through a ball screw–nut assembly is shown in *Figure 12.10*. Ball screws with internal and external ball recirculation are obtainable with efficiencies of over 90%. The external method is less expensive than the internal one, but a larger nut is necessary. To increase the load-carrying capacity the number and area of contact points between the nut and screw needs to be increased; this is done by using roller screws.

Backlash can be eliminated in screw–nut systems by adjusting preload through the use of shims. This is done by splitting the nut at right angles to the screw and inserting a shim, i.e. a packing piece, between the nut halves. The nut is then clamped together, e.g. by using cap screws or rivets. If the shim has been accurately ground, no significant increase in friction will occur.

Other transmission systems are available, but conventional gear train systems and screw–nut systems are the major ones used for machine tools. Industrial robots utilise a wider range of transmission elements, e.g. synchronous belts, chains, pretensioned metal bands, pulleys, and harmonic and cycloidal gear drives, which provide extremely high speed-reduction ratios in a compact package.

12.5 Fluid power actuators

The use of hydraulic drives in machine tools has decreased considerably. This has been due to various factors such as high noise levels from power packs, vibrations transmitted through the fluid from the pump, leaks, and analogue control systems using sensitive servovalves with which few maintenance engineers are familiar. However, although hydraulic drives are no longer common on machine tools, they do still have some advantages over electric drives, i.e. they can withstand high shock loads, hydraulic force can be applied directly where required without the use of mechanical transmission elements, and they can be used in fire-risk areas as the voltages required by the servovalves are very small. For these reasons there are still industrial applications for the simpler hydraulic actuators, e.g. some robots employ linear and rotary hydraulic actuators.

For simple, high-speed, inexpensive movement, pneumatic actuators are still extremely popular. Being intrinsically safe they can be used in high fire-risk areas and explosive atmospheres. They are seldom used in servocontrolled applications or for moving heavy loads. This is due to the compressibility of air which prevents reliable positioning and speed control. Pneumatic power systems and devices are ubiquitous in industry and are well known by almost every factory worker.

Since hydraulic and pneumatic actuators are both fluid power devices, their basic construction is similar. The main differences lie in the wall thicknesses and material, as hydraulic systems have to operate at internal pressures of around 20 MPa, whereas pneumatic devices operate at about 0.7 MPa.

Hydraulic and pneumatic motors are not considered here because they are seldom, if ever, used on modern machine tools. Pneumatic motors are used for driving small tools such as automatic screw drivers, and hand-held deburring tools. For heavy-duty work, hydraulic torque amplifiers are available; these are, in effect, servocontrolled hydraulic motors. Pneumatic linear actuators are available in the normal ram-type-cylinder form or as 'rodless' cylinders. Hydraulic linear actuators are available as simple cylinders or as 'linear amplifier' units which incorporate an electric pilot motor, a servovalve, and a hydraulic ram. *Figures 12.11* to *12.13* illustrate the configuration and method of operation of the fluid power cylinder, pneumatic rodless cylinder, and hydraulic linear amplifier. Fluid power rotary actuators are normally of the rack and pinion or rotary vane type; their method of operation is shown in *Figures 12.14* and *12.15*.

12.6 Fluid power actuator control valves

Located between the power source (i.e. a compressor or a pump for pneumatics or hydraulics, respectively) and the actuators, are the control valves. Their function is to direct the energy contained in the fluid to the appropriate parts of the system.

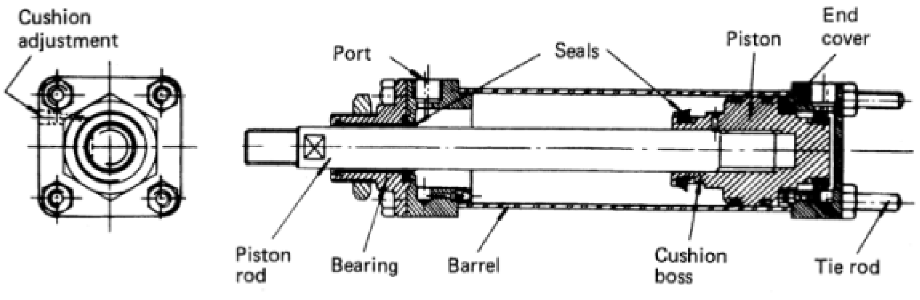


Figure 12.11 Double acting pneumatic cylinder. (Courtesy of Norgren Martonair)

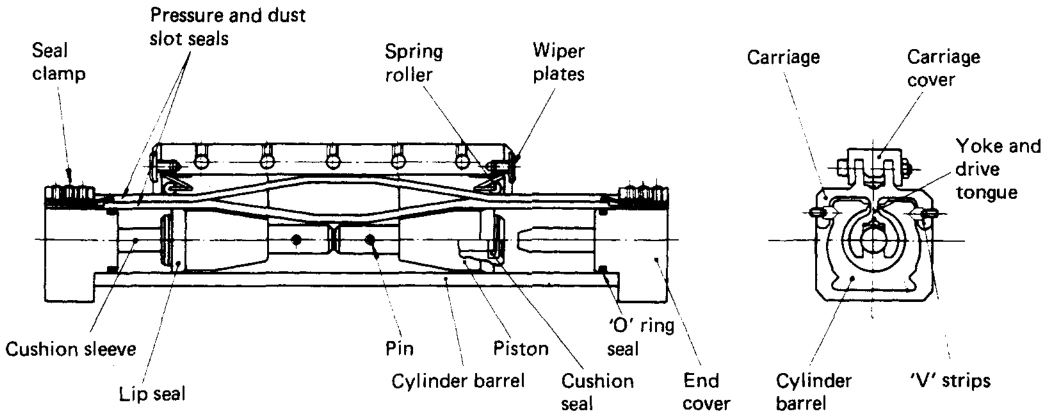


Figure 12.12 Double acting pneumatic rodless cylinder. (Courtesy of Norgren Martonair)

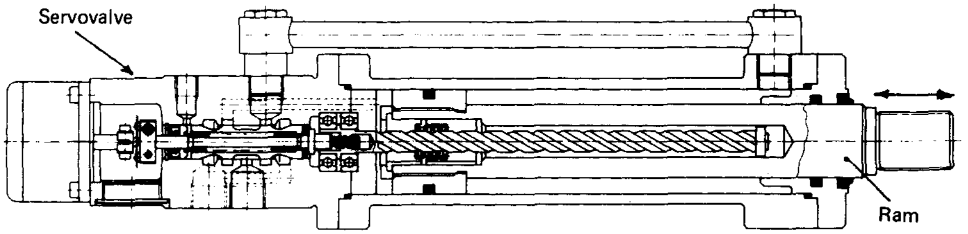


Figure 12.13 Hydraulic linear amplifier

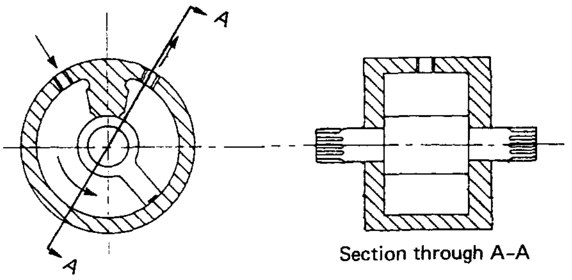


Figure 12.14 Fluid power rotary vane actuator

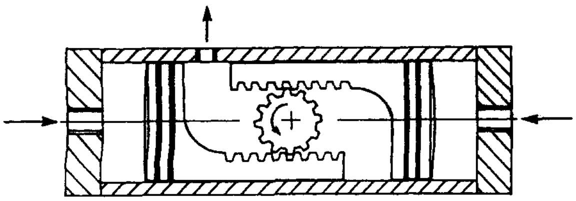


Figure 12.15 Fluid power rack and pinion rotary actuator

12.6.1 Control-valve specification

For fluid power control valves the International Standard ISO 1219 prescribes the symbols to be used to represent these valves. Examples of these are shown in *Figure 12.16*.

In order to fully define a fluid power control valve, a number of factors must be specified.

(1) The number of ports on the valve. These are the main ports necessary for the control of the air to and from the actuators; the number does not include the pilot signal ports.
(2) The number of positions of the internal switching element. This is either two or three, the former being more common in pneumatic circuits and the latter in hydraulics. The number of ports and the number of positions available for a valve are represented by stating the numbers with a solidus separating them, e.g. '5/2 spool valve' represents a five-port, two-position valve.
(3) The control method used to cause the internal switching. Methods include pneumatic, electric, or some form of mechanical action such as a spring, a lever, or a manual push button. Their equivalent symbols are shown in *Figure 12.17*.
(4) The rest position, or state, of the valve, i.e. whether normally closed (NC) or normally open (NO).
(5) Whether the valve is monostable or bistable. A monostable valve has only one stable position. This is often the rest position in the absence of a control signal, it is

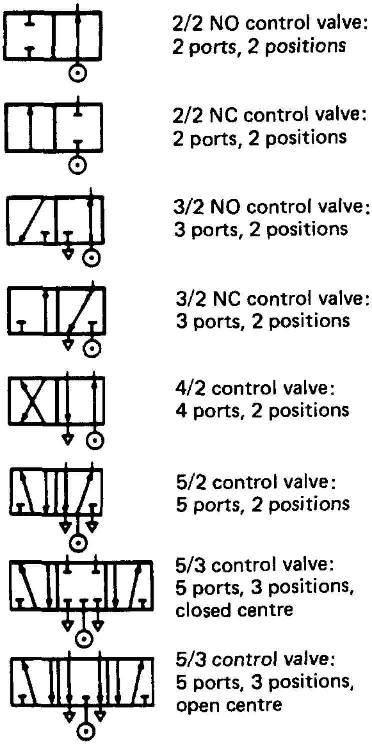


Figure 12.16 Fluid power control valve symbols

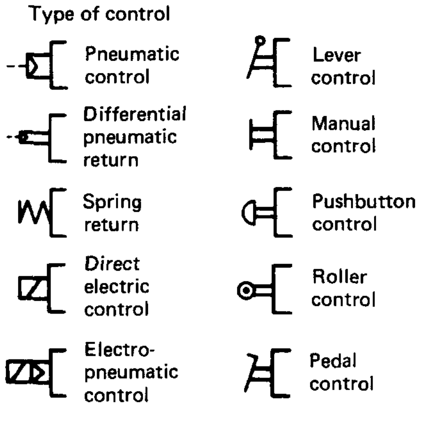


Figure 12.17 Control method symbols

usually achieved by the action of a spring or differential air pressure. The solenoid operated poppet valve described later is a monostable device. A bistable valve remains in the last position attained whether or not a command signal is present. A command signal is required for every change of state. Because of its nature, a monostable device has, in effect, a memory function, i.e. it 'remembers' the last command. The spool valve, also described later, is an example of a device that is normally supplied in a bistable form, although monostable versions are available.

12.6.2 Pneumatic control valves

In pneumatic systems the flow of air to the actuators is controlled by pilot and spool valves. These are non-servo systems, similar valves are also used in non-servo hydraulic systems. Pilot valves receive the control signal, either mechanical or electrical, and translate this into a low pressure air pulse to the spool valve. Pilot valves are usually of the poppet type. These have an internal switching element which opens or blocks the movement of air by moving a seal off of, or on to, a seat. The switching element may be operated mechanically, but in modern control systems solenoid valves are common. A solenoid valve takes an electrical signal and converts it into a pneumatic signal pulse to a spool valve. The solenoid coil is energised by a d.c. or an a.c. current which is switched to it by the system controller, e.g. a programmable logic controller (PLC). When this happens the core is caused to rise, thus opening the air passage to the spool valve. A spring is usually incorporated to return the valve to its rest position after the coil has been de-energised. Poppet valves are available in two rest position states, i.e. normally closed (NC), and normally open (NO).

Solenoid operated pilot valves are often mounted on the spool valve as illustrated in *Figure 12.18*. The type of spool valve shown is a five-port linear slide valve, i.e. it has a high pressure input port, two outlet ports to an actuator, and two exhaust ports.

The air pulse from the pilot valve alters the position of the switching element in the spool valve, thus directing high pressure air to the appropriate actuator chamber. The switching element is a shouldered shaft or 'spool', and movement of this within the valve chamber allows either high-pressure air to flow to the actuator, or exhaust air from the actuator to flow to atmosphere. A simple pneumatic system is shown schematically in *Figure 12.19*; this incorporates a double-acting pneumatic cylinder, a spool valve, and two mechanically operated pilot valves. This circuit provides the piston with a reciprocating action.

Other elements are also available. For example, the shuttle valve is used to direct flow when two alternative air signals are to be used in a system. It is a very simple bistable device allowing air to flow only in the direction determined by the last command signal. Flow regulators are another example; these are used to restrict air flow and can be used on the exhaust ports of actuators to control speed of movement. Conversely, the speed of actuator movement can be increased by use of

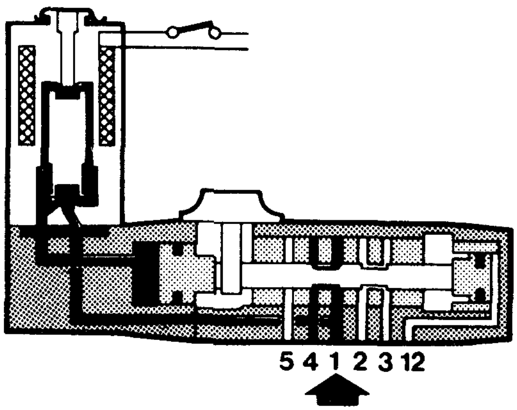


Figure 12.18 Solenoid operated pilot valve. (Courtesy of Joucomatic/Techno-Nathan International)

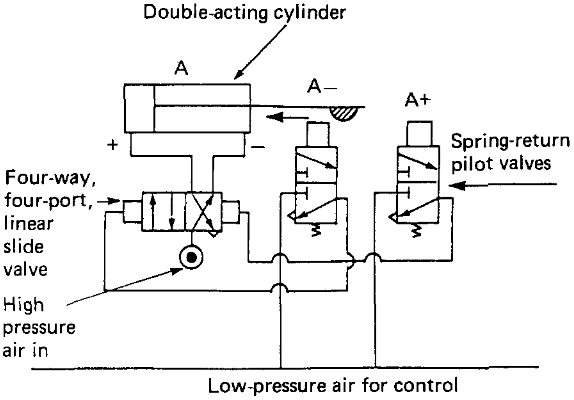


Figure 12.19 A pneumatic circuit producing a reciprocating action

quick exhaust valves fitted directly to the actuator exhaust ports. There are also electropneumatic and pneumatic–electric interface modules available. These convert electric signals to pneumatic and pneumatic signals to electric, respectively, thus allowing simple interfacing to electronic control systems.

12.6.3 The hydraulic servovalve

All the valves mentioned above operate in either the fully on or the fully off state. In hydraulic systems, where the fluid used is incompressible, it is possible closely to control the actuators by controlling the fluid flow. This closed-loop servo control of force, displacement, and speed is effected by the hydraulic servovalve. The type described here indicates the general principle of operation, although it is only one of the many types available.

The valve shown in *Figure 12.20* comprises three main elements: a torque motor, a pilot stage, and a spool-valve assembly. These elements are integrated into a single compact

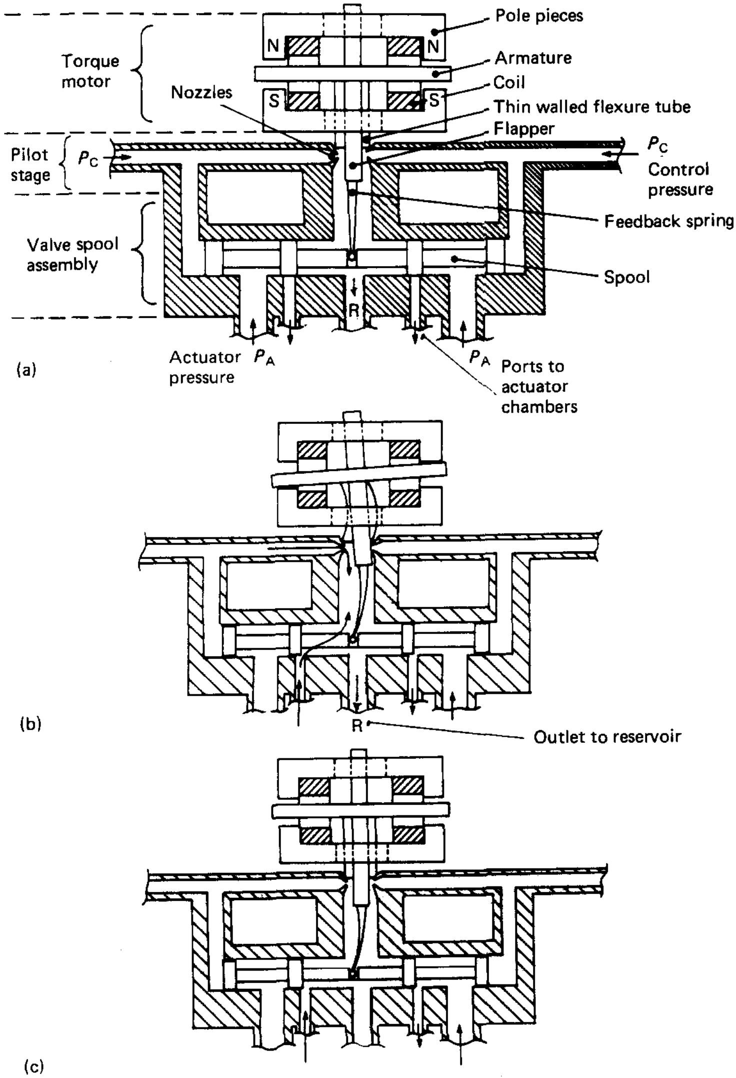


Figure 12.20 Principle of operation of a Moog flow control servovalve

unit in the actual product. The valve's function is analogous to that of the spool valve discussed in the previous section, with the exception that the present one allows fine control of the fluid flow to the actuator. Here, a flapper amplifier stage is used; other valve types have spool, deflector jet, and jet pipe amplifiers. In the figure, control fluid P_c enters the valve at the right- and left-hand sides and initially flows freely through the valve to the return port R. The high pressure fluid P_a required to operate the actuator can be seen entering the spool chamber through two ports at the base of the valve. Ports to the actuator chambers are shown adjacent.

The main elements of the first stage, i.e. the torque motor, are the upper and lower pole pieces, two permanent magnets, an armature and a coil. Attached to the armature is the flapper and feedback spring. These are surrounded by a thin-walled flexure tube which prevents hydraulic fluid from entering the motor chamber. The permanent magnets polarise the pole pieces, so creating a magnetic field across the gap in which the armature is supported. A d.c. current proportional to the command signal from the system controller is passed through the coil. The current flowing in the coil creates a torque on the armature which causes it to rotate through a small angle. This torque is controlled by the d.c. current.

On experiencing torque the armature causes the flapper to block the flow of fluid P_c through one of the nozzles in the second stage. For example, if the armature experiences an anti-clockwise rotation then the flapper will be pushed against the right-hand nozzle, this will increase the pressure on the right-hand side of the spool, pushing it to the left. A path is now opened in the valve's third stage for the high-pressure fluid to flow from the pump, through the servovalve, to the appropriate actuator chamber. Fluid in the low-pressure side of the actuator can also now flow back through the valve to the hydraulic reservoir.

As is evident from the figure, when the spool moves to the left it will also move the feedback spring to the left. This produces a clockwise torque on the armature, thus tending to pull the flapper away from the nozzle again. When this 'feedback torque' equals the torque created by the current flow, the flapper will return to its central position and the spool will remain in its new position as P_c will be equal on each side. Thus, assuming constant pressure, the flow of hydraulic fluid to the actuator will be proportional to spool position and the spool position will be proportional to the input current.

Servovalves are used in conjunction with feedback transducers to provide full closed-loop control, e.g. a potentiometer for position servoing, a tachometer for velocity servoing, and a load cell containing a strain gauge for force servoing. *Figure 12.21* shows a basic system.

12.7 Feedback transducers

The previous sections in this chapter have covered the principles of operation of a machine tool control system, the actuators necessary to effect movement, and the methods of control of these actuators. This section examines the elements necessary to provide feedback to the system controller on position, proximity, displacement, velocity, and force. There is a wide range of these feedback sensors and, therefore, only the most common ones are covered here.

12.7.1 Position and proximity sensors

12.7.1.1 Limit switches

These are widely used as simple position feedback devices. Their main component is usually a 'microswitch'. This is a small switch that requires only a very slight movement to cause a mechanical snap action which opens or closes the electrical connection. The casing and means of activating the microswitch are available in a number of forms, e.g. plunger, roller, lever, and roller lever (see *Figure 12.22*).

The switches are located at the desired limits of travel of the machine-tool elements. They often act as safety devices, being fixed at a point just beyond the expected software-set stopping point of a machine tool table or robot arm. A mechanical trip or 'dog' contacts the switch to complete, or open, an electric circuit and shut off the machine tool. These switches can be

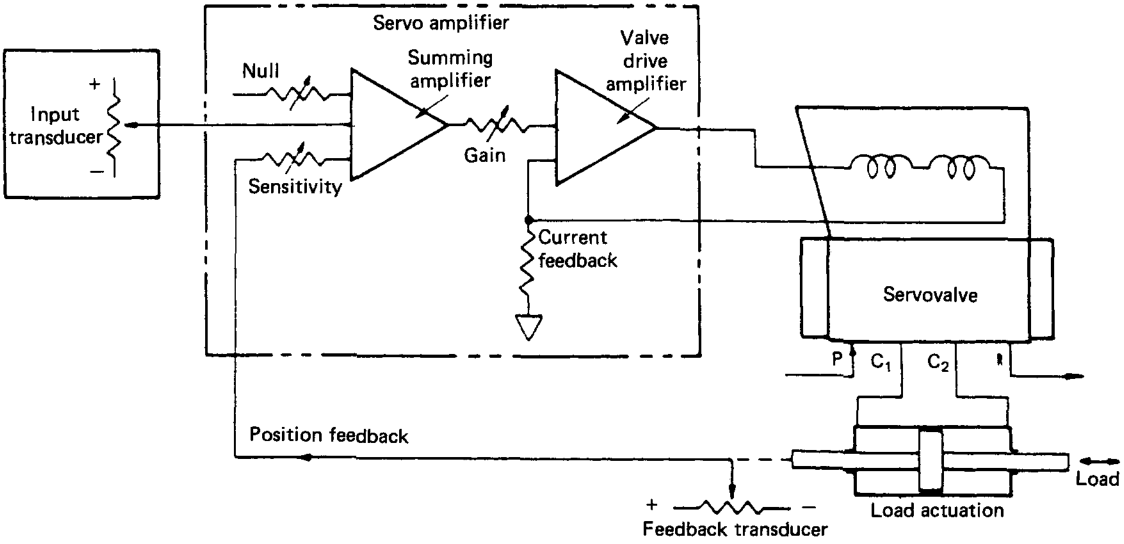


Figure 12.21 A typical hydraulic servovalve positioning system. (Courtesy of Moog Inc.)

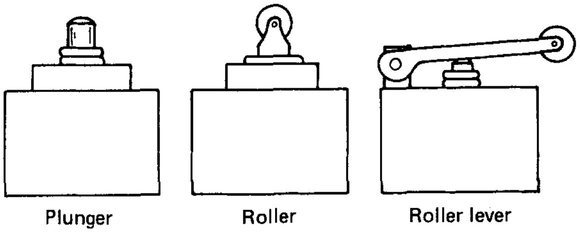


Figure 12.22 Some limit-switch configurations

used in a normally open or a normally closed mode. Use outside the limits of the machine tool is also common, e.g. they are incorporated in safety barriers or interlocked gates, and are used as component detection or pallet identification devices.

Microswitches have also been used for sequence control. For example, a linear array of microswitches can be linked to various fluid power solenoid valves. The switches can then be switched on or off in the required sequence by setting pegs on a rotating drum to contact the microswitches in the appropriate order, i.e. the pattern of pegs will determine the sequence, and the speed of drum rotation will determine the sequence cycle time.

12.7.1.2 Reed switches and magnets

These are usually found on pneumatic cylinders where their function is to detect the position of the piston within the cylinder chamber. The switch component is mounted on the outside cylinder wall at the position at which the piston is to be sensed. The casing contains a ferromagnetic flexible blade, or 'reed', type switch. The piston within the cylinder contains a permanent magnet. The cylinder itself is non-magnetic, being made of aluminium, non-magnetic stainless steel, or brass. As the piston comes within close proximity of the switch, the magnetic field from the magnet causes the contacts to close. An electric circuit is thus completed, allowing a control signal to pass to the controller (see *Figure 12.23*). These devices can enable detection of the piston to close tolerances, and some have a light-emitting diode (LED) fitted to signal closure of the switch. They have the advantage of being non-contacting devices, thus providing greater reliability. Being small they provide a compact position detection system without the need for additional sensors such as limit switches, which occupy often much needed space.

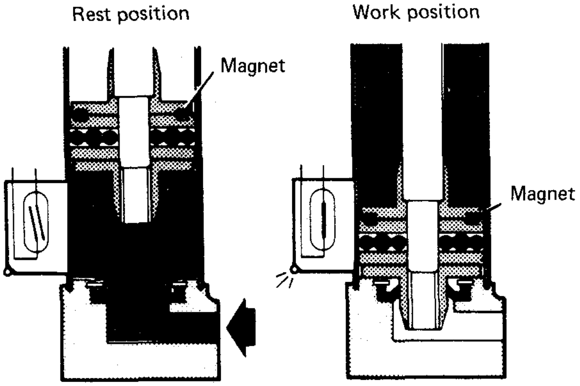


Figure 12.23 Reed switch and magnet used on a pneumatic cylinder. (Courtesy of Joucomatic/Techno-Nathan International)

12.7.1.3 Inductive sensors

These are used for non-contact proximity sensing of metallic machine elements or other objects. Often called 'magnetic pick-ups' they appear in the form of a coil wound round a permanent magnet and enclosed in a small cylindrical case (see *Figure 12.24*). The control system is connected to the coil in such a way as to enable the voltage across the coil to be measured. Initially the magnet produces a steady magnetic field around the case and no voltage is induced in the coil. However, as a metallic object approaches the pick-up, so the magnetic field around the coil is changed, thus causing a corresponding change in the induced voltage in the coil. A threshold value can be set in the controller to determine at what voltage level the object should be deemed to be in adequate proximity to the pick-up. It is possible, over very short distances, to calibrate the pick-up in such a manner as to determine the distance of the object very accurately by monitoring the induced voltage.

12.7.1.4 Capacitive sensors

These are used for non-contact proximity sensing of non-metallic objects. They can be used in a safety system around the machine tool to sense the presence of personnel, or in non-contact keypads where the proximity of a finger or hand would be sufficient to operate a switch. An oscillator-type circuit is shown in *Figure 12.25*. Here a 'capacitor' is formed comprising the metal plate, the atmosphere which acts as the dielectric, and the non-metallic object which acts as the other

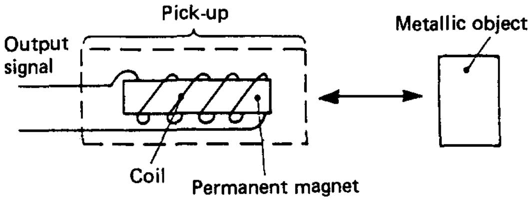


Figure 12.24 The inductive sensor

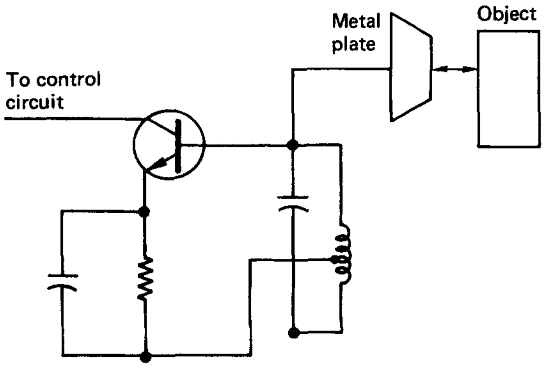


Figure 12.25 The capacitive sensor

'plate'. When the non-metallic object is not present, circuit oscillation is maintained by the feedback, compensating for losses occurring in the coil and capacitor circuit. The circuit is so adjusted that if the circuit load is increased the oscillations stop, so causing a current change which can be used to operate a switch. The presence of the non-metallic object causes the necessary additional load to stop the oscillations.

12.7.1.5 *Photosensors*

These are used for non-contact sensing in many situations, e.g. for detecting the presence of a part in an automatic machining or assembly system, for counting components passing along a conveyor belt, or for constructing light curtains in safety barriers. They are also widely used in optical displacement and velocity feedback elements. Photocells are available that are sensitive to light in the visible or infra-red spectrum. The most common type is the 'photoconductive' or 'photoresistive' sensor, this is composed of materials the electrical resistance of which is decreased when exposed to light.

In the cadmium sulphide type (*Figure 12.26*), a layer of cadmium sulphide is sandwiched between two metal electrodes (alternatively, a layer of lead selenide can be used for infra-red sensing). The top electrode can be made as a thin wire mesh of tin oxide in order to allow light to reach the photoconductive layer. If the cell is in total darkness and a voltage is applied across the cell, the cell provides a resistance of several megaohms, and a negligible current will flow. If light then impinges on the cell the resistance drops to a few ohms and a relatively large current will flow. Since the resistance varies inversely with the light intensity, the light level can be determined by measuring the resistance.

Another type of photocell utilises the photosensitive properties of many semiconductors such as diodes and transistors (see *Figure 12.27*). Again when a voltage is applied while the photocell is exposed to light, a current proportional to the light intensity will flow.

Both the above-mentioned photocells require the application of a voltage from an external source. A different type is the photovoltaic sensor. Commonly known as a 'solar cell' this produces a voltage when exposed to light. Photovoltaic materials used are selenium, silicon and germanium. As in the inductive sensor mentioned earlier, the photocell can be used as a simple ON/OFF device by setting a threshold value, or it can be used as an analogue device. In its ON/OFF form, light is often supplied from a LED from which light passes in a beam to the cell. When the beam is broken by a person or component the change in voltage causes a signal to be passed to the controller. The cell may also be used to register the presence of a reflective or bright surface against a dark background. The main problem with the use of light sensors in the industrial environment is that the cell can be occluded by contaminants such as oil or dust.

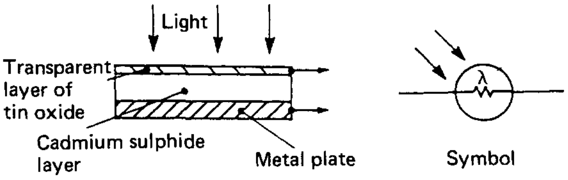


Figure 12.26 Cadmium sulphide photocell

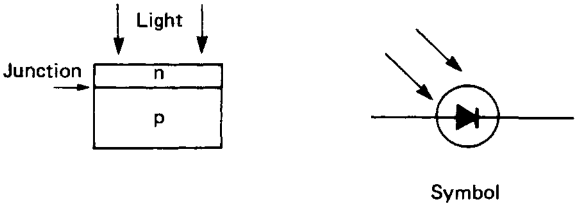


Figure 12.27 Photodiode photocell; n, electron rich material; p, electron deficient material

12.7.1.6 *Pneumatic sensors*

Sometimes called 'air-jet sensors', these are used for non-contact sensing in areas where contaminants prohibit the use of photocells. They are suitable for use in high-temperature, corrosive, and explosive atmospheres and in situations where the object to be detected has an imprecise location. Two types are common: the back-pressure sensor and the interruptible-jet sensor.

Figure 12.28 shows the principle of operation of the back-pressure sensor. A tube-shaped probe is supplied by air at constant pressure P_c, this supplies air to the control chamber via a fixed orifice. The pressure of the air in the control chamber is determined by the distance between the end of the probe and the object to be detected. As the object approaches the sensor the escape of air decreases and the pressure P_c increases. In fact, when the distance x is less than $0.25D$ the pressure P_c varies proportionally with distance. Air loss and sensitivity to strong currents usually limit the diameter D to a maximum of about 4 mm, which limits the dynamic range to about 1 mm. Thus the device can be used as a simple proximity detector over distances of about 1–15 mm, but can be used as an analogue distance measuring device at distances under 1 mm.

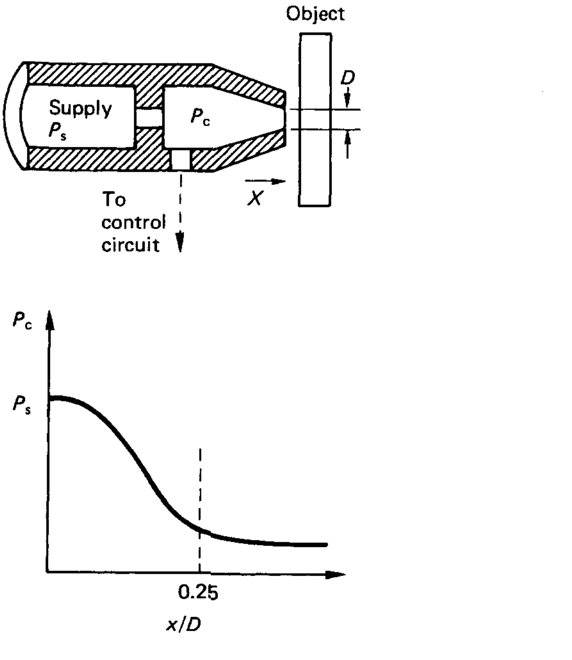


Figure 12.28 Pneumatic back-pressure proximity sensor

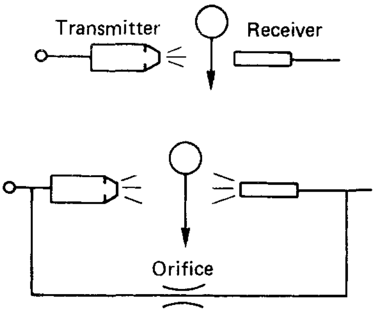


Figure 12.29 Pneumatic interruptible-jet proximity sensor

The interruptible-jet sensor can detect objects passing between the emitter and the detector. It is analogous in operation to the LED and photocell arrangement. *Figure 12.29* shows the principle of operation. A jet of air from a low-pressure air source is projected across a gap to a receiving orifice. The air pressure within the orifice is monitored. If an object now passes through the gap and breaks the jet, the pressure in the receiver drops and this is noted by the controller. For use in dusty environments, the receiver can be pressurised as shown to prevent the possibility of clogging. Supply pressures range from 0.07 to 0.2 bar. Typically the transmitter may have a diameter of around 3 mm and the receiver a diameter of about 6 mm. Combined transmitter and receiver units have detection distances of around 6 mm, and separate units can detect up to 100 mm.

12.7.2 Displacement transducers

These may be linear or rotary, analogue or digital, and incremental or absolute. They can all be used to provide direct or indirect feedback. Direct feedback occurs when the transducer is mounted integrally with the elements being monitored. For example, if the movement of a machine-tool table is being measured relative to a fixed bed, then a feedback transducer mounted on the table and/or the bed will provide a direct measurement of displacement. However, if the transducer is mounted on, say, the shaft of an electric drive motor, then the feedback will be indirect because the movement being measured must still be transmitted to the machine-tool table. The latter method is less precise as mechanical inaccuracies and backlash may be present in the transmission system. Direct systems are, therefore, to be preferred, although there are many cases where they cannot be used, e.g. 'revolute' geometry industrial robots have a design that prevents practical direct position feedback of their end effector.

Linear transducers are commonly used on machine-tool tables and slideways. They provide reliable and precise direct displacement feedback using optical gratings and encoders. Linear potentiometers can be used over short distances, but are not used on machine-tool axes. Rotary transducers in the form of optical shaft encoders, potentiometers and resolvers are popular for mounting integral with servomotor assemblies for indirect displacement feedback.

Analogue transducers produce a continuously changing signal analogous to the value being measured; potentiometers and resolvers are analogue transducers. Digital transducers produce digital signals, e.g. an optical grating will produce, after signal conditioning, a train of square-wave pulses for input to an electronic counter.

Incremental transducers provide signals that must be added as they arrive at the counter to determine the displacement from a known starting, or 'zero', point. However, with an absolute transducer each signal pattern specifies uniquely where the device is. Thus with an incremental transducer, when powering up the system, or restoring power after a failure, the machine axes must be returned to the 'zero' point, but with an absolute encoder this is not necessary. The following describes the operation of some position transducers.

12.7.2.1 *Potentiometers*

These are available in both linear and rotary form, the latter probably being the most popular. They are analogue devices with a voltage output proportional to the displacement of a wiper blade in contact with a resistive element. Their principle of operation is shown in *Figure 12.30*. A d.c. voltage V is applied across the resistive element R. The resulting voltage v across the distance r from the positive terminal to the wiper blade is measured. In the case of a rotary potentiometer this voltage is proportional to the angle of the wiper blade relative to the datum; in the case of a linear potentiometer it is simply proportional to the distance travelled. The voltage v is equal to Vr/R. It is possible to derive a 'voltage constant' for the potentiometer. If we let the voltage constant be K, then $v = Kr$, where K is the voltage constant of the potentiometer in volts per radian for rotary potentiometers and volts per millimetre for linear potentiometers. The value of the measured voltage therefore indicates displacement, and direction is indicated by increasing or decreasing voltage values. In practice, the shaft of a rotary potentiometer may be connected to the element being monitored either directly, through a gear train, via a rack and pinion, or by a synchronous belt.

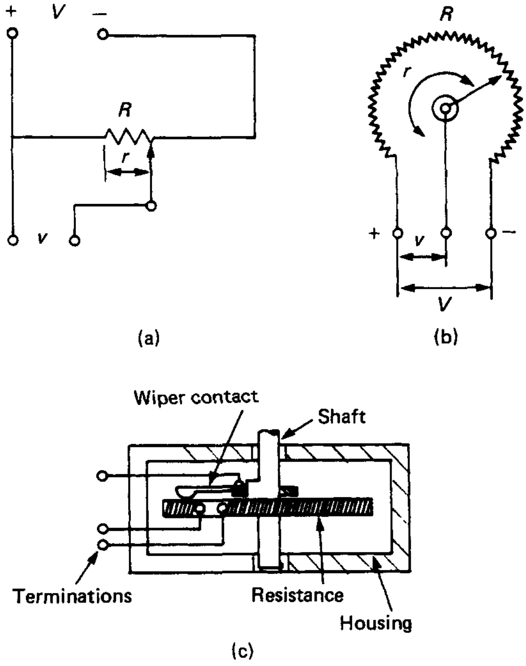


Figure 12.30 The potentiometer: (a) circuit; (b) layout; (c) construction

12.7.2.2 Resolvers

Again these are analogue devices, but in this case the input is an a.c. voltage. Resolvers are essentially rotating transformers with an output proportional to the angle of a rotating element (the rotor) with respect to a fixed element (the stator). Physically, they resemble small electric motors. Typically, there is a single winding on the rotor and two on the stator. The electrical connections to the rotor are made by slip rings and brushes. The principle of operation is shown in *Figure 12.31*. The two stator coils are 90° apart and, when the rotor is excited by a signal $V \sin(\omega t)$ the voltage induced across the stator coils is:

$$v_1 = V \sin(\omega t) \sin \theta$$

and

$$v_2 = V \sin(\omega t) \cos \theta$$

where θ is the angle of the rotor with respect to the stator.

In practice, the stator coils may be fed with an input voltage and an output voltage induced in the rotor coil. For example, the stator coils may be supplied with two square-wave voltages 'sin' and 'cos' at a frequency of about 2 kHz, their amplitudes will be the same but they will be 90° out of phase with each other. The voltage induced in the rotor coil can be filtered and conditioned to produce a square-wave voltage signal the phase position of which relative to, say, the 'sin' voltage is a measure of the angular position of the rotor relative to the stator.

It is also possible to get multipole resolvers. The difference in operation occurs in the relationship of the phase shifts to the physical angle of rotation. For example, if in a two-pole resolver an electrical rotational angle of 360° corresponds to a physical rotation of 360°, in a 10-pole resolver the same physical rotation will produce an electrical rotational angle of 5 × 360°.

The rotor shaft can be connected to the motor shaft or other elements in the same way as the potentiometer. The output signal is connected to an electronic counter which records both the number of full revolutions of the rotor and the relative angular displacement at any point in time.

Another device similar to the resolver is the *synchro*. This is similar in construction to the resolver except that it has three coils wound on the stator, each 120° apart. Again, when the rotor coil is energised, voltages are induced in the stator coils the amplitude and phase of which are indicators of the angular displacement of the rotor relative to the stator.

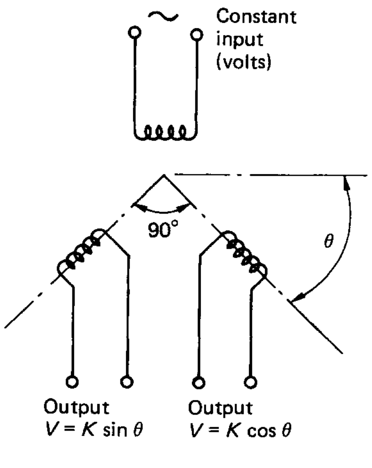


Figure 12.31 The resolver

12.7.2.3 Inductosyns

These are not normally found on new machines and are now mostly sold for replacement. They can be obtained in both linear and rotary forms. The rotary type can be considered as a flat, many-poled, synchro. For example, an inductosyn with 2000 poles would have an electrical rotational angle of 360° for a physical rotation of the rotor of 0.36°. In linear inductosyns a voltage is induced in the windings in the long linear scale analogous to a rotor, with the slider and its windings being analogous to the synchro stator.

12.7.2.4 Optical encoders

These are photoelectric devices obtainable in linear and rotary configurations. They operate by using a photosensor to monitor changing light intensity created by optical gratings. A grating comprises opaque lines and transparent spaces of equal width. These gratings are commonly produced by depositing very thin chrome layers onto a glass substrate using thin-film processes such as vacuum deposition or sputtering. Alternatively, the gratings may be etched onto stainless-steel strips so that reflected light can be used. These strips are bonded onto rigid steel bars which have a similar coefficient of expansion to that of the machine slide. It is also possible to etch the gratings directly onto a solid stainless-steel bar. Reflected light systems avoid the problems inherent in using glass scales, i.e. different coefficients of expansion to the machine elements and sensitivity to contamination. With linear devices, grating periods of 20, 10 and 8 μm, and with rotary devices up to 9000 lines per revolution, are typical.

Method of operation Two types of encoder exist, i.e. the incremental and the absolute. The following applies to the incremental type, the absolute type is described later. *Figure 12.32* shows the principle of operation. The encoder is basically composed of two parts; the graduated glass scale which may be either straight line or circular, and the scanning unit. This unit comprises the light source, a collimating lens, a scanning reticle with its own gratings, and the photocells which often are of the photovoltaic type. In linear encoders the scale is usually fixed to the machine bed and the scanning unit fixed to the machine table. In the rotary types the scanning unit is held relatively fixed in the encoder housing, while the optical disc is connected to a rotating shaft which is coupled to a motor drive shaft, leadscrew, etc. (see *Figure 12.33*).

There are three configurations of scanning unit. In one configuration the lines in the reticle and scale are parallel. In this case the light intensity falling on the cells increases or decreases, as the lines either occlude the light or coincide with each other, so forming a space for it to pass, as shown in *Figure 12.34*.

In the second configuration the graduations are magnified by the use of moiré fringes. This effect is produced when two lined patterns are superimposed with one pattern of lines slightly inclined to the other (see *Figure 12.35*). When illuminated by either transmitted or reflected light, fringes are created which lie approximately at right angles to the graduations. The 'wavelength' or spacing of these fringes is directly related to θ, the angle of inclination between the lines, i.e. the 'wavelength' increases as the angle decreases. In practice, the scanning unit contains the reticle with the lines inclined to those on the fixed grating. As the unit passes along the grating the fringes move vertically up or down, depending on the direction of horizontal travel. For each horizontal movement of the unit by one grating, the fringe moves vertically by one 'wavelength'. Thus, if a mask with a narrow horizontal slot is placed over the reticle, then the intensity of light passing

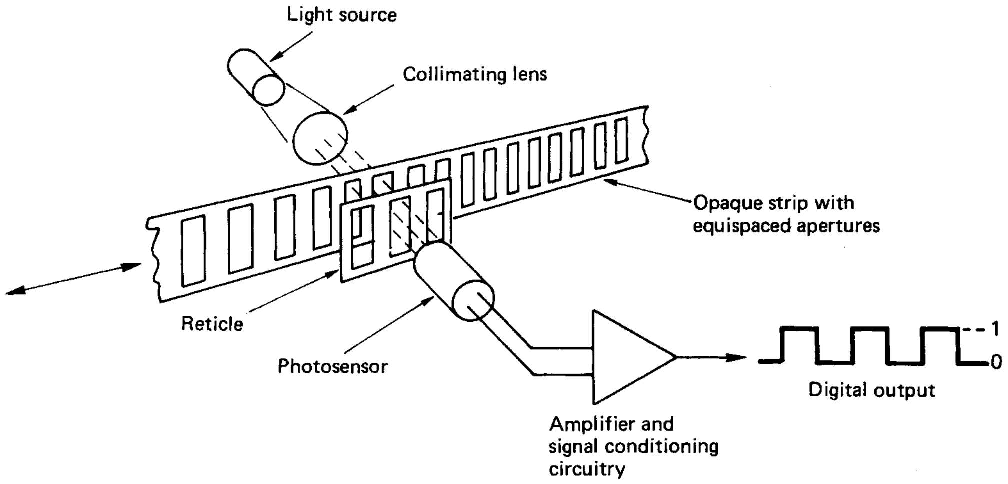


Figure 12.32 Principle of operation of an optical encoder

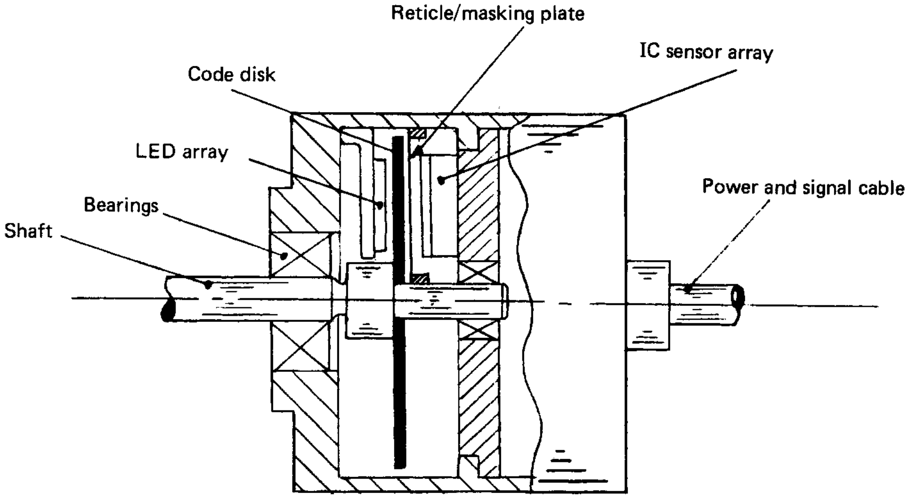


Figure 12.33 Section through an optical shaft encoder

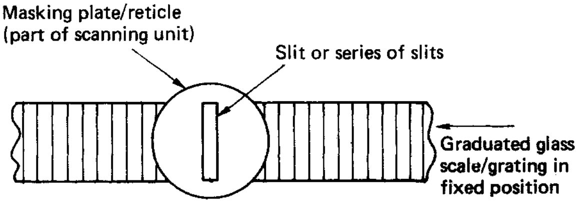


Figure 12.34 Optical encoder arrangement using parallel slits

through the slot will wax and wane as the fringes pass under it. The distance between the fringes is much greater than that between the gratings. This means that a number of slots can be used to monitor the movement of the fringe over one 'wavelength'; for example, four slots could be used with four photosensors to provide four signal pulses for one grating pitch movement.

The third method uses the interferential scanning principle for obtaining extremely fine resolutions and high accuracies. For example, the Heidenhain LIP 101R incremental linear encoder has a grating period of 8 μm, a signal period of 4 μm, and the interpolation is capable of giving measuring steps of 0.01 μm, although under practical conditions maximum accuracies of ± 0.2 μm could be expected. The encoder uses the reflected-light method. It consists of a solid steel rectangular bar onto which a grating has been etched, and a scanning unit containing the reticle and light source. Briefly, the system operates as follows. The grating period is too fine to allow the use of moiré fringes and, therefore, the interference principle is used to produce light and dark bands which can be observed by the photocells. When light waves of the same wavelength are in phase they combine to increase the resultant amplitude,

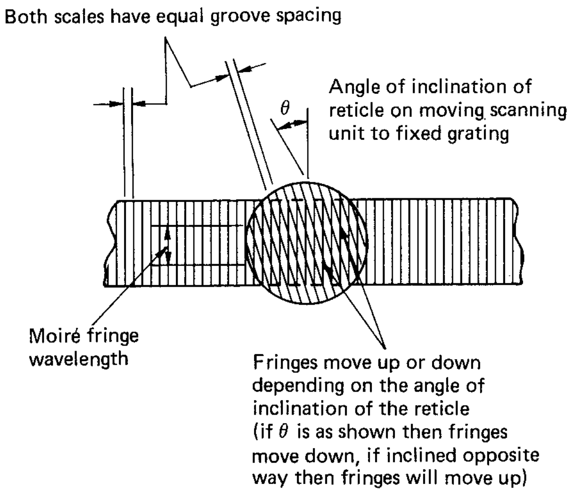


Figure 12.35 Optical encoder arrangement using moiré fringes

and hence light intensity, to a maximum. When they are out of phase by 180°, i.e. half a wavelength, the resultant amplitude which is the algebraic sum of both wave amplitudes, is at a minimum. If the waves are of the same amplitude and 180° out of phase then the algebraic sum of the amplitudes is zero, interference occurs and darkness is observed.

Light of the same wavelength can be produced from a monochromatic light source or by diffraction, i.e. splitting the light into its constituent elements by passing it through a narrow slot. In the interferential scanning principle, as collimated light passes through the gratings in the reticle most of its goes straight through with zero diffraction. However, some light is diffracted to the right and left of the grating in various degrees. Due to the nature of this diffraction and the fixed phase relationship that the waves initially have to each other, it is possible to observe interference patterns that are created as the phase relationships change due to the relative movement of the reticle. It is these interference patterns that are scanned by the photosensors to provide the displacement measurement.

Signal processing Irrespective of which of the three methods of operation is employed, a light signal fluctuating in intensity is produced. This light signal causes a photocell to generate a voltage in the form of a sine wave corresponding to the increase and decrease in light value. Interpolation and digitising circuitry takes this sinusoidal signal and transforms it into a square wave pulse train. The sinusoidal signal is first amplified, then interpolated to provide measuring steps of from 1/20 to 1/200 of the grating period. Finally, using Schmitt triggers, the signal is changed into a square wave. The resulting pulse train can then be used as direct input to the counting circuitry of the machine controller or display unit.

The system can therefore count the number of pulses from a datum to determine displacement; using an internal clock, the system can use the displacement to derive velocity and acceleration. In incremental encoders, an additional grating is usually employed to act as a local zero datum. In order to determine the direction of travel the signal phase relationships are usually examined. For example, in a rotary encoder two tracks and two sensors can be employed (see *Figure 12.36*). These tracks generate signals with a 90° phase shift, which is equivalent to a quarter-period time displacement. Either of

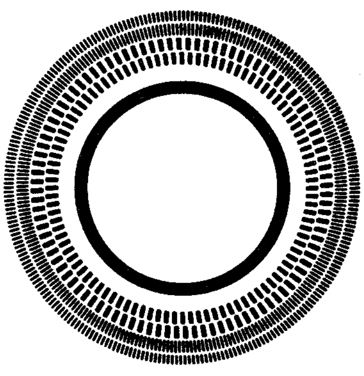


Figure 12.36 Incremental encoder disk with two pairs of tracks. (Courtesy of Gaebridge)

the tracks can be used to provide the displacement data, but it is the comparison of their signals to determine the relative phase which provides direction information. It is also possible simply to use two sensors over one track. These sensors are positioned exactly one-half segment width apart. Again the signals produced by the sensors will be phase shifted by 90° relative to each other and, by noting which sensor responds first, the control system can determine the direction of movement.

On incremental shaft encoder discs the gratings can be arranged to provide direct angular feedback, e.g. 360 or 3600 gratings per revolution, or chosen in combination with a mechanical gearing system such that metric or imperial readouts can be obtained. For example, if the gearing is arranged so that one revolution of the disc corresponds to a mechanical movement of, say, a machine table of 1 in., then a disc with two sets of tracks could be used. One set would contain 1000 gratings per revolution, and the other 2540, thus allowing signals corresponding to a resolution of 0.001 in. and 0.01 mm to be selected, as desired.

Absolute encoders Though available in linear and rotary types, absolute encoders are more often used in the latter form. Two problems with incremental encoders are:

(1) they only produce output signals while power is applied and the slideway is moving or shaft is rotating; and
(2) if power loss occurs the actual position reading is also lost, which means that on powering up again the encoders must be brought back to a datum or 'home' position from which to resume counting.

Absolute encoders overcome these problems by allowing the position of, say, a shaft to be uniquely defined immediately on power up, whether or not there is movement. The absolute encoder is a more complex device than the incremental type and may cost twice as much. It has a large number of concentric tracks with one sensor per track (see *Figure 12.37*). A twelve track encoder, for example, provides 4096 uniquely defined positions for each revolution. Resolution can be increased by increasing the number of tracks and by using multiturn encoders.

It is possible to use a pure binary coded disc, but this means that all digits on, for example, a four-track disc must change simultaneously from 15 to 16—this can produce ambiguity problems which may result in the recording of incorrect information. Methods of overcoming this include using antiambiguity tracks on the disc, or tightening up mechanical

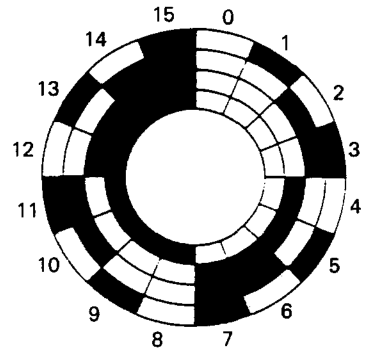
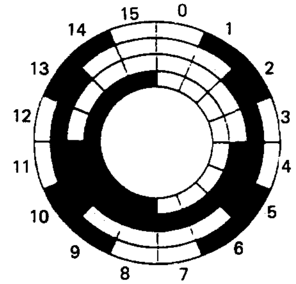


Figure 12.37 Scheme of an absolute disk using binary code



Decimal	Natural binary	Gray code
0	0000	0000
1	0001	0001
2	0010	0011
3	0011	0010
4	0100	0110
5	0101	0111
6	0110	0101
7	0111	0100
8	1000	1100
9	1001	1101
10	1010	1111
11	1011	1110
12	1100	1010
13	1101	1011
14	1110	1001
15	1111	1000

Figure 12.38 Scheme of an absolute disk using Gray code and a table showing a comparison of codes

tolerances. The most common method, however, is to use Gray coding (see *Figure 12.38*) in which only one digit changes at any moment. Conversion to true binary, or other code, if required, can be done using electronics within the encoder itself.

Encoders in operation Linear encoders are normally mounted directly onto the bed and table of a machine tool, Rotary encoders are indirect in that they are connected to the motor drive, leadscrew, rack and pinion, or other gearing, usually by using collet or sleeve and grub screw arrangements. All types of encoder are available in a wide range of physical sizes. They operate from 5 or 12 V power supplies with loadings of around 60 or 120 mA, depending on type. Interfacing capability is usually provided, e.g. RS232 and RS422, with some encoders incorporating their own microprocessor systems for error checking, etc.

12.8 Conclusion

This chapter has described the individual control elements used in machine tools. For very high-volume production runs some machines use 'hard' automation. Examples of this are cam-controlled multispindle lathes where the cams control the sequence and movement of the cutting tools, and dedicated assembly machines where cams, in conjunction with pneumatic elements, control the sequence and movement of pick-and-place units, presses, inspection probes, etc. At the other end of the spectrum, manual machines used for single-unit production such as prototyping, employ only the basic elements of drive unit, gearing system and, probably, leadscrew. An electronic aid to control may involve the use of a linear encoder linked to a digital readout. These readouts give easily read feedback on tool position and can provide direct conversion from metric to imperial scales if required. Fully programmable numerically controlled machine tools take the output from the feedback devices and use this as input to the control unit. It is these numerically controlled machine tools that have been the main concern of the chapter.

Further reading

BARBER, A., *Pneumatic Handbook*, 7th edn, Trade and Technical Press (1989)

BRADLEY, D. A. *et al.*, *Mechatronics*, Chapman and Hall, London (1991)

BRINDLEY, K., *Sensors and Transducers*, Butterworth-Heinemann, Oxford (1988)

CRISPIN, A. J., *Programmable Logic Controllers and Their Engineering Applications*, McGraw-Hill, New York (1990)

DOTE. Y and KIROSHITA, S., *Brushless Servomotors*, Oxford University Press, Oxford (1989)

ELECTRA-CRAFT CORPORATION, *DC Motors, Speed Control, and Servo Systems*, Electra-Craft Corporation (1980)

GAYAKWAD and SOKOLOFF, *Analogue and Digital Control Systems*, Prentice-Hall, New York (1988)

HUGHES, A., *Electric Motors and Drives*, Butterworth-Heinemann, Oxford (1990)

JACOB, J. M., *Industrial Control Electronics*, Prentice-Hall, New York (1988)

KENJO, T., *Stepping Motors And Their Microprocessor Controls*, Oxford University Press, Oxford (1985)

MAIR, G. M., *Industrial Robotics*, Prentice-Hall, New York (1988)

MORRIS, N., *Control Engineering*, McGraw-Hill, New York (1991)

PARR, A., *Hydraulics and Pneumatics—A Technician and Engineer's Guide*, Butterworth-Heinemann, Oxford (1991)

PARR, E. A., *Industrial Control Handbook, Vol. 1 Transducers, and Vol. 2 Techniques*, Blackwell Scientific, Oxford (1987)

SINCLAIR, I. R., *Sensors and Transducers—A Guide for Technicians*, 2nd edn, Butterworth-Heinemann, Oxford (1992)

13 Communication and Integration Systems

G Miller

B Etter

Contents

13.1 Computer architecture 13/3
- 13.1.1 Basic computer architecture 13/3
- 13.1.2 Computer architecture classification 13/4
- 13.1.3 Computer size classification 13/4

13.2 Computer operating systems 13/5
- 13.2.1 Operating system techniques 13/5
- 13.2.2 Addressing and memory management schemes 13/6
- 13.2.3 Commonly used operating systems 13/7

13.3 Computer languages 13/7

13.4 Peripheral devices 13/8
- 13.4.1 Classification of peripheral devices 13/8
- 13.4.2 Storage devices 13/8
- 13.4.3 Semiconductor storage 13/8
- 13.4.4 Magnetic storage devices 13/9
- 13.4.5 Hard copy output devices 13/10

13.5 Human–computer interfaces 13/10
- 13.5.1 Computer output interface devices 13/10
- 13.5.2 Computer input interface devices 13/11

13.6 Networks 13/17
- 13.6.1 Basic communications classification 13/17
- 13.6.2 Transmission media 13/20
- 13.6.3 Network topologies 13/21
- 13.6.4 Network protocols 13/21
- 13.6.5 Typical networks 13/22

13.7 Databases 13/23

13.8 Knowledge-based systems 13/24

This chapter deals with the hardware and software requirements of computer-aided and computer-integrated manufacturing systems. Such topics as computer architecture, operating systems, languages, peripheral devices, man–machine interfaces, networks, databases, and expert systems are described. The incorporation of computer technology into a manufacturing environment is a complicated and expensive proposition, and should be thoroughly planned before implementation. However, the use of computers in a manufacturing and industrial environment can improve productivity through increased work throughput, increased efficiency, and better quality control. Further information regarding each of these subjects can be found in the publications listed in the Further Reading section at the end of this chapter. In order to understand the information regarding computer systems and communications hardware and software, a glossary of commonly used computer terms is given in *Table 13.1*.

Table 13.1 Glossary of terms commonly used in computing

Term	*Description*
Address	Logical location of information in a memory system
ALU	Arithmetic logic unit: functional part of the CPU which performs arithmetic operations on data
Architecture	Physical structure of a computing system
Array	Multidimensional group of numbers; multiple vectors
Bit	Most basic form of information in a digital system with discrete values of 0 (OFF) and 1 (ON)
Bus	Series of parallel logic channels used to transmit information
Byte	Pattern of 8 bits
CCD	Charge coupled device: a secondary NDRO, volatile, storage device
CPU	Central processing unit: main computational/controlling device in a computer system
Data	Information used or modified by instructions
DRO	Destructive readout device: device whose contents are destroyed when read
Duplex	Method of data transmission where information is transmitted in two directions simultaneously
GPIB	General purpose interface bus: class of data links used for communication with instrumentation
I/O	Input/output: devices used for input or output
Instruction	Information controlling the movement or alteration of data
LAN	Local area network
MODEM	Modulator/demodulator
MBM	Magnetic bubble memory: a secondary NDRO, non-volatile, storage device
Non-volatile	Storage device which loses information after power is lost
NDRO	Non-destructive readout: device whose contents are not destroyed when read
Program	Series of instructions ordered to achieve a certain result
RAM	Random access memory: memory which is accessed in the same time cycle, independent of location
ROM	Read only memory: memory which is accessed only during read operations
RWM	Read write memory: memory which is accessed for reading or writing
Register	High speed storage device normally located within the CPU
Simplex	Method of data transmission where information is transmitted in one direction only
Vector	One-dimensional group of numbers
Volatile	Storage device which loses information after power is lost
Word	Multiple groups of bits, the size of which depends on the computer system

13.1 Computer architecture

13.1.1 Basic computer architecture

Figure 13.1 illustrates the basic components of a modern computer system. The ALU is the component of the CPU which actually performs computation on the data presented to

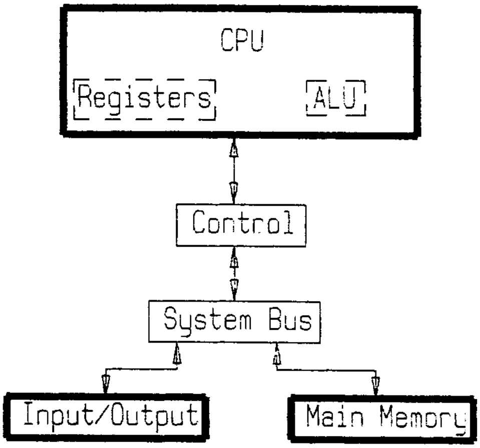


Figure 13.1 Basic components of a computer system

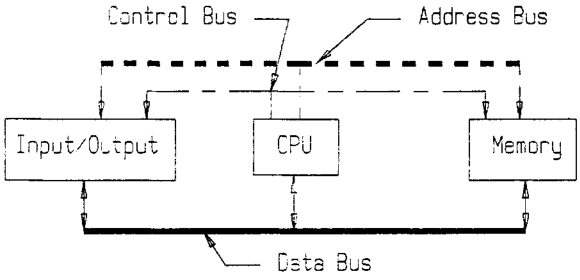


Figure 12.3 Computer system bus structure

the CPU via the system bus. The control logic decodes the incoming instructions and is responsible for synchronisation in the CPU and with the system bus. The register's function is to facilitate the movement of data and instructions within the CPU. The system bus is the physical connection between the CPU and memory and external I/O devices.

The basic structure of the system busses of a common computing architecture is illustrated in *Figure 13.2*. Here an additional I/O processor is used to control operations with slower mechanical devices. This additional processor allows the host CPU to perform computational tasks independent of the I/O processes. The control bus contains the instructions to be performed between the processors and memory. The address bus contains the address of the memory location being accessed by the controller. Data coming into or out of the processors are sent via the data bus.

13.1.2 Computer architecture classification

The flow of information to and from the CPU is commonly referred to as a 'stream'. The architecture of modern computers can be divided into four basic groups based on the instruction and data streams.

The first class of computers is the simplest, where the CPU executes one instruction with one segment of data. These machines are classified as single instruction stream, single data stream (SISD). The instruction and data flow are illustrated in *Figure 13.3*. Most microcomputers and many minicomputers fall into this category. Most of the DEC PDP-11 series minicomputers portray this type of architecture.

The next step in the hierarchy is a single instruction stream, multiple data stream (SIMD) computer. This type of machine normally includes one program control unit and multiple execution units. The data and instruction paths are shown in *Figure 13.4*. Most of the machines with this type of architecture are minicomputers. Among these are the ILLIAC IV and the Goodyear STARAN.

The complement of the above architecture is the multiple instruction stream, single data stream (MISD) machines. Pipeline processing is the predominant architecture in this class. This technique is illustrated in *Figure 13.5*. The data flow through the processor in a linear, pipeline fashion. This pipeline is divided into m segments. Each segment may be the operand of a separate instruction. This allows a single processor to act as an m-unit processor in many applications. Most of these machines are classified as mainframes or supercomputers. Examples of this type of architecture include the CRAY-1, CDC STAR 100 and the TI ASC.

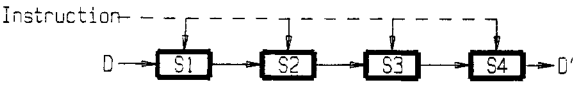


Figure 13.3 SISD instruction and data flow

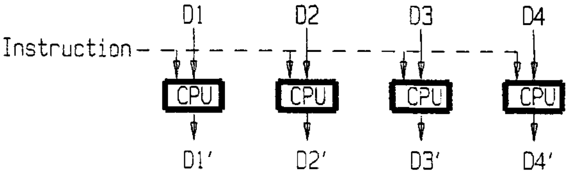


Figure 13.4 SIMD instruction and data flow

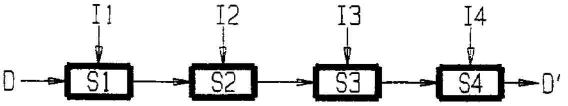


Figure 13.5 MISD instruction and data flow

The last category is that of multiple instruction stream multiple data stream architectures. This category includes machines capable of executing independent programs simultaneously. Many experts refer to this architecture as the class of 'multiprocessors'. This structure is illustrated in *Figure 13.6*. These computers may also be classified as vector computers with the inclusion of vector operations in the machine instruction sets. Machines with this type of architecture are expensive and are normally reserved for applications where computational speed is critical. This type of architecture is normally achieved with multiple SIMD or MISD architecture computers linked together.

13.1.3 Computer size classification

The four classes of computer architecture can be implemented in almost any size of computer system. The classification of system size is normally based on a system's computational speed and intended usage. Most multiuser systems are minicomputers and mainframes; however, a multiuser operating system can be implemented on microcomputers. A workstation is normally distinguished as a high-power single-user system; however, many minicomputers are implemented in a workstation configuration. Whether implemented in a single or a multiuser environment, workstations are normally used for dedicated processes which are computationally extensive, e.g. graphics applications.

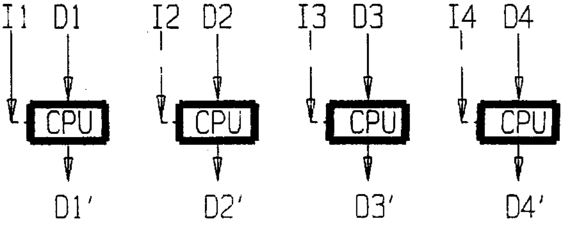


Figure 13.6 MIMD instruction and data flow

The smallest size modern digital computer is classified as a microcomputer. Most of these systems are miniaturised versions of the larger computer systems. Microcomputers have the following general architectural charcteristics:

(1) data and memory busses are normally 8 or 16 bits wide;
(2) ALU operations are normally register–register or register–memory operations;
(3) all data types are either integer or decimal; and
(4) interrupts are used to transfer control to peripheral devices.

One of the major limiting factors in microcomputers is the smaller bus width. Memory and I/O access are slowed down due to this physical limitation. Recent developments have increased the physical address bus to 32 bits, making the abilities of this class of computer comparable to many larger systems. Other physical-size constraints prevent the use of microcomputers in larger system applications.

Minicomputers are distinguished from microcomputers primarily by the physical address size capabilities. Most modern minicomputers have address widths of 32 bits or more. This permits the use of faster bus access without complex addressing schemes. The larger bus widths also provide for an increase in computational accuracy due to the increase in significant digits (bits) in each execution cycle. This class of computer also provides for an increase in data types. This permits single-cycle execution for many operations which require multiple executions in a smaller system. The resulting increased instruction-set size frequently exceeds 200 commands.

The class of mainframe computers, or supercomputers, is distinguished from the smaller systems primarily by computational speed and price. Several minicomputers have similar architectures to the larger mainframes, only the implementation in the large computer maximises all available hardware technologies to provide a more optimal system. Another common distinction is an increased width, which increases computational accuracy and speed. The increased bus width also provides for more addressing structures and data types, which increases the size of the basic instruction set.

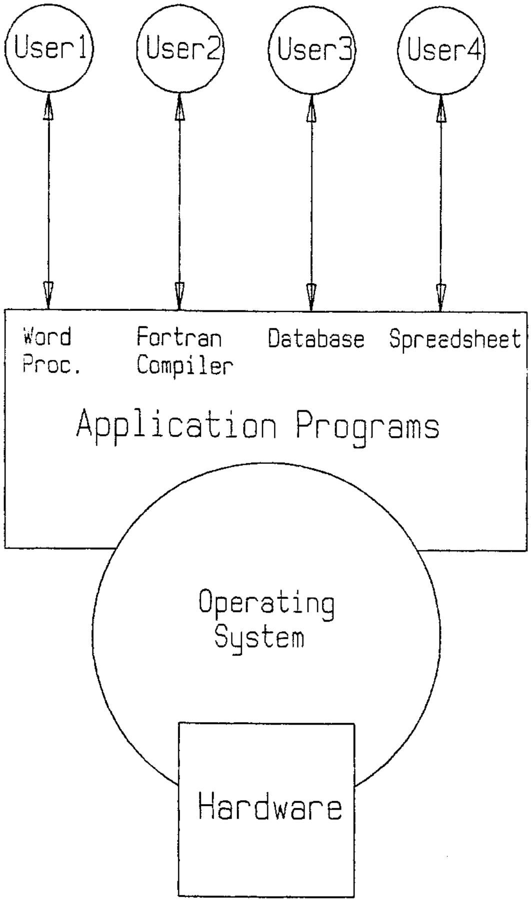


Figure 13.7 Operating system relationships

13.2 Computer operating systems

A computer operating system is the software (programs and data) that controls the hardware and applications software of a computer so that they can constitute a useful system for carrying out calculations. From a computer user's perspective, an operating system provides a means of accessing and controlling applications software. From a computer system manager's perspective, an operating system controls the central processor, various internal and external memories, various peripheral devices, and physical communications hardware in order to allow the computer as a whole to be optimally managed and operated. Operating systems are used to control the number of users of a computer system, the amount of computer time and memory allocated to each user, the interaction between the computer user and peripheral devices, and the access to applications software systems. *Figure 13.7* depicts the relationship between the computer operating system and the hardware, applications software, and computer user.

13.2.1 Operating system techniques

The most basic operating system is a single-user system which executes one program at a time. This type of system is most prevalent in dedicated systems such as microcomputers or workstations. Single-user systems can be used with larger computers; however, such an implementation would not use the computer to its maximum efficiency.

Batch systems are used in multiuser computer applications. Each individual's job is processed in sequential order of submittal: first in, first out (FIFO). This operating system can cause major bottlenecks in system performance when time-consuming jobs are being done.

A modification of the batch operating system is the multiprogramming system. This mode divides the CPU execution time among several different submitted programs, in order of priority and sequence. Programs that exceed a predetermined execution time are temporarily suspended to allow the next program to operate for the same allotted amount of CPU time. The overhead of switching between programs can become excessive, depending on the architecture of the system and the amount of CPU time allotted. This method also aids CPU operations when interacting with slower devices in I/O operations. It should be noted that batch systems and multiprogramming systems would perform faster on a system with a parallel architecture; however, multiprogramming does not imply true simultaneous processing.

The principle of multiprogramming can be developed to provide for an interactive multiuser system. This technique is

classified as 'time sharing'. Most modern computer systems are capable of time-sharing functions. Each user interacts with the system as though it were a dedicated computer. This is the predominant operating system technique currently used in multiuser computer systems. Many systems use a combination of batch and time sharing to provide optimal flexibility and performance.

A variation of the single-user system is multitasking, where a single user uses several programs 'simultaneously'. The operating system actually uses a time-sharing technique for each of the enacted programs. The CPU time allotments are such that the user is unaware of the switching between the various programs (tasks).

13.2.2 Addressing and memory management schemes

The various operating systems schemes can rapidly deplete the available amount of main memory for most computer architectures. The operating system must provide for access to data in secondary memory storage. This secondary storage is commonly a magnetic-medium device, or other high-storage device. These devices have slower access times; therefore access by the primary CPU(s) should be minimised.

Most computer programs exhibit a high degree of structure. Current trends in computer science encourage the development of structured programming. Programs may be segmented into modules based on this structure for ease of transfer from secondary memory devices. Instructions and operands are transferred to and from the secondary and main memories for operation by the primary CPU(s). These segments vary in length based on the amount of main memory available and the length of the program module. Although segmentation corresponds to program module limits, it results in a great deal of fragmentation of main memory. A complex memory allocation operating system is required to avoid this fragmentation in segmentation methods.

A similar process known as 'paging' transfers fixed blocks of data to and from secondary memory. Paging avoids the problem of main memory fragmentation because fixed blocks of memory are moved; however, the page boundaries normally do not coincide with program boundaries.

A virtual memory addressing scheme comprises paging, segmentation, or both, to provide the user with an apparent infinite memory area. There are many different variations of this memory allocation scheme, but *Figure 13.8* illustrates the basic differences between segmentation, paging, and virtual memory. Certain programs may require such frequent swapping of memory that the system spends more time moving memory than executing the program. This is known as 'thrashing'. The detection and prevention of thrashing has resulted in rather complex virtual memory operating systems.

Computer architectures with multiple processors frequently access memory at the same time. The process of interleaving enables two or more processors to access memory by partitioning the memory addresses into separate modules, as indicated in *Figure 13.9*. If two or more processors attempt to access the same module at the same time, a contention occurs. The interleaving process aids in access of most memory address; however, branching instructions in the program may cause departures from the normal process.

The cache memory is a small section of memory located between the processor and the main memory. This technique is similar to paging; however, the cache is between the main memory and the processor rather than between the secondary and main memories. The primary difference is that cache memories are even faster than the main memory, similar to the dedicated CPU registers. There is a risk of thrashing due to the similarity to virtual memory systems. The control of a

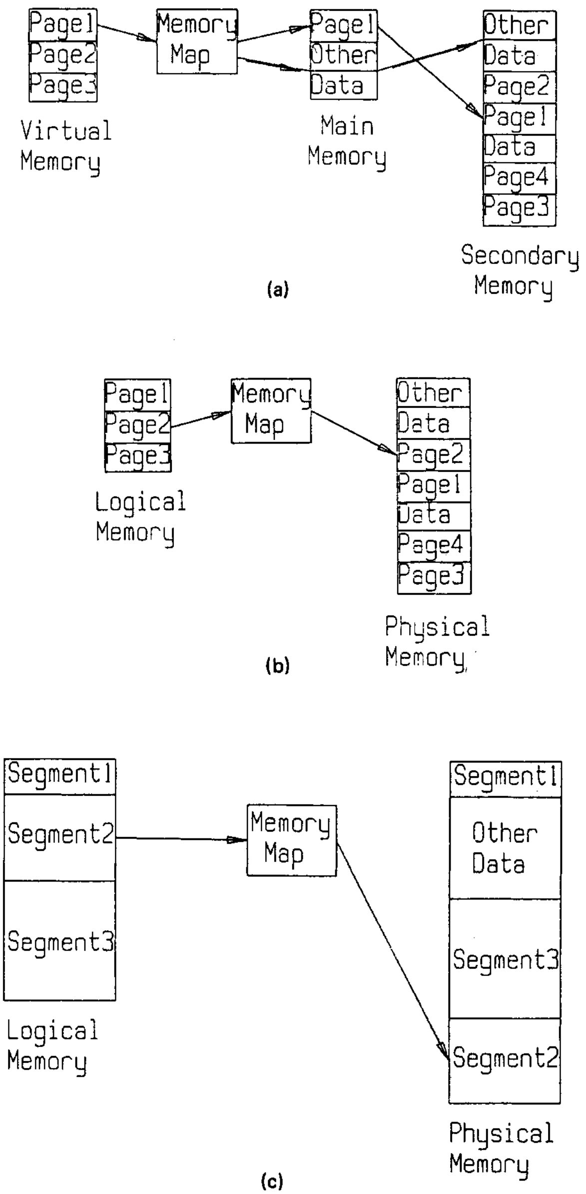


Figure 13.8 (a) Virtual memory. (b) Memory paging. (c) Memory segmentation

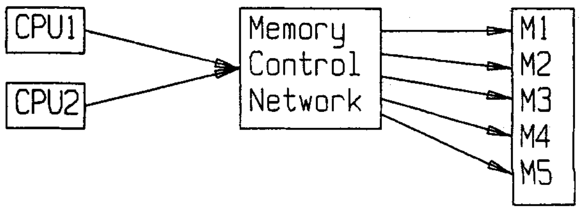


Figure 13.9 Memory interleaving

virtual memory system is handled with operating system software, whereas the control for cache is processed with dedicated logic circuits. The size of the pages is much smaller with the cache than with the virtual memory system. The primary purpose of the cache is to increase system speed by making the memory access time similar to the main processor speed.

13.2.3 Commonly used operating systems

Unix is a general-purpose time-sharing system which was originally developed at Bell Laboratories for the PDP 11/20 computer.[1] This system has blossomed into widespread use throughout the world on a variety of computers. However, this system is primarily used on the PDP/11, VAX 11/780, Amdahl 470, Univac 1100, Interdata 7/32 and 8/32, and various microcomputer systems.

This system is usually written in the C language and entails approximately 8000 lines of code with another 800 lines of assembly language code, making this system quite small compared with most other operating systems. The user interface is terse, clean and relatively easy to use. Installation on a new computer system is relatively easy and the system is portable to other compatible computer systems with ease.

Unix supports multiple processes. The CPU scheduling gives priority to I/O bound jobs. However, priorities are continually recomputed every second, and the system defaults to a round-robin time sharing control if no I/O commands are present. The key Unix features include:

(1) a programmable command language interpreter,
(2) the ability to control asynchronous processes,
(3) compatible file, device, and interprocess I/O,
(4) a hierarchical file system with demountable volumes, and
(5) a high degree of portability.

IBM has developed the *OS/360* operating system for its 360 and 370 class of computers.[2] The original intent was to develop a single operating system which could be used for a wide variety of computers (size and function) marketed by IBM. OS/360 is actually a series of operating systems which provide a vast array of functions and controls and support a wide range of applications programs. Thus, the OS/360 operating system is very large and very expensive. The system comprises over 200 000 lines of code.

OS/360 provides support for many higher level computing languages such as Cobol, Fortran, and PL/1, as well as support for assembly language code. The job control language is very sophisticated and detailed. The file control system is also very detailed and contains a large number of utility programs to control user-generated files. This system supports interactive programming and contains specialised subsystems to support various user-specific areas such as business databases. Due to the complexity of the OS/360 operating system, the interaction between software and hardware must be specified at the time of initial installation. The OS/360 system is not portable between computers and is not easy to use. However, due to its large number of features and the vast number of IBM systems in use today, the OS/360 system is widely used.

There are two basic versions of this system: OS/MFT uses fixed memory regions, and OS/MVT uses variable memory regions. The MVT version allows several user jobs to run concurrently, with each job assigned a section of computer memory. This system does not support paging. Thus, the memory initially assigned to a job remains with the job until its conclusion. With the advent of the IBM 370 architecture, virtual memory was added to the MFT version of OS/360. Two versions of MFT were subsequently developed to employ virtual memory: VS1 creates one large virtual address space, and VS2 (now called MVS) provides each user with his own virtual memory.

The *1100 Executive* system is employed by the Univac 1100 series of computers. This system supports various batch and interactive languages including Fortran and Basic. This system provides good support for data communications and time-sharing operations.

The *MCP* operating system supports the Burroughs B5000, B6000 and B7000 series of computers and is primarily intended for business applications including support of COBOL. The operating system is written in Algol and provides excellent support for large databases and high-level programming.

The *TOPS* operating system was written to support the DEC system 10 and system 20 computers. This system is primarily designed for scientific applications involving the Fortran language. Networking and time sharing are both supported.The *VMS* operating system has been written for the various VAX computer systems and contains features that are similar to that of TOPS.

The *DOS* (disk operating system) system is an industry standard for microcomputer operating control. This system is primarily intended for single-user, single-job applications, although the latest versions have attempted to address the expanding field of multifunction operation, which has been developed for the 386 series of microcomputers. The DOS system supports a wide variety of applications programs in several languages, including Fortran, Basic and Pascal.

13.3 Computer languages

C is a general-purpose programming language featuring economy of expression, modern control flow, data-structure capabilities, and a variety of operators and data types. It is best known as the primary language of the Unix operating system, although it is used in many other environments and on a vast array of computer models and sizes, from mainframe to microcomputer. The C language has been used with numerical, scientific, text-processing and database-management applications. It is relatively easy to use. The system is very portable and is not tied to any one particular application area.

Fortran was one of the first high-level languages developed for general use. The name is an abbreviation for 'formula translator'. Fortran is primarily intended for scientific programming and it supports various data types. Fortran was one of the first computer languages to provide a structured, yet user programmable, system for handling data input and output, complex and double precision operations, and control of data-storage allocation. It is a memory-intensive language that is not easily transportable between computer systems. The latest version, Fortran 77, is designed to support structured programming, which previous versions did not allow. The WATFIV version of Fortran contains excellent error diagnostics and rapid program compilation. Fortran is available for virtually all types and sizes of computer, from microcomputers to mainframes. Because it was one of the original scientific programming languages, Fortran is still widely used throughout the world.

Cobol is another of the older, high-level computer languages; it was developed during the late 1950s, near the time of Fortran development. Cobol is primarily designed for business applications and is an abbreviation for 'common business oriented language'. The original aim of Cobol was to be used on a variety of machines without difficulty. Thus, user training and program installation is relatively simple. A Cobol program is split into four divisions: identification, environment, data, and procedure. The 'identification' division merely

states the title of the application.The 'environment' division links the data and procedure to a specific type of computer. The 'data' division describes the files and records which will be employed by the program. The 'procedure' division specifies and controls the processing and computation which will be accomplished. Cobol, like Fortran, has been so widely used, that it has now been developed as an ANSI standard.

Algol is a successor to Fortran and is an abbreviation for 'algorithmic language'. It was developed as a method of developing algorithms that is machine independent. Its major features include a flexible, yet powerful, method of handling loop calculations and the ability to combine a set of program statements into a single compound statement. Algol can handle large amounts of data efficiently by applying dynamic data-storage allocation, which assigns a location for data at the time that it is needed in a calculation. This storage is then released after the calculation is completed. Algol has been widely used in in Europe, although its use in the USA has been limited. Recently, Pascal has supplanted Algol as the most widely used algorithmic language in Europe.

Pascal is a successor to Algol which has been designed for structured programming. Pascal is an easy language to learn and use and is very portable. However, it does not contain many of the features of other algorithmic languages, including dynamic data storage. In addition, the number of predefined functions within Pascal is less than that of other languages. However, Pascal allows the programmer to define various types of data, which makes this a unique language in that regard.

PL/1 was developed for the IBM 360 computers and draws on many features found in Fortran, Cobol and Algol. It is widely used for IBM based mainframe systems and supports both scientific and business applications. However, because it contains the features of many other languages, it is a large, memory-intensive language that is not easy to use or learn.

APL is a procedures-oriented language that is usually used in a real-time interactive mode by a user. A key feature of APL is the ability to use single-statement, multiple-function operations that would require the use of numerous statements in other languages. Data are structured in arrays and calculating functions have no specific hierachy. The use of APL can often generate very concise, yet powerful, programs. However, input and output control is limited, since the system is primarily designed for use at an interactive terminal. Use of APL has not been widespread, because it has not been used for a wide variety of applications.

Basic is an abbreviation for 'beginners all purpose symbolic instruction code'. It was designed as a simple, easy to use, easy to learn, interactive computer language. It has been widely used by beginners and is primarily employed on microcomputers. Basic has severe limitations in data storage and data manipulation and is primarily designed for use in small programs. Basic employs relatively simple user-defined functions, subroutines and matrix operations, and was originally conceived as a simpler, more rapid version of Fortran. Unlike Fortran, Basic lacks widespread control of input and output functions. However, Basic can be used as a rapid, interactive controller to access other programs written in various languages, and for various applications.

13.4 Peripheral devices

13.4.1 Classification of peripheral devices

Peripheral devices are normally perceived to be devices located outside (peripheral) the main computer, although they are actually distinguished by function, not location. Peripheral devices may be located within the computer housing itself, e.g. disk devices and secondary memory. These devices are distinguished by a slower access time; however, they offer a much lower cost of storage of information.

13.4.2 Storage devices

Most of the attention regarding peripheral devices is focused on storage devices. Most of these devices may be used for both input and output; however, some are dedicated for input. The information stored in these devices is performed before the user accesses the media. There are two primary factors involved in the selection of storage media: cost and access time. *Figure 13.10* illustrates the trade-off in access speed and cost per bit of some typical storage devices.

Most main memory is classified as random access memory (RAM) where the processor can access any address, independent of physical location. Information accessed in a serial storage device must be accessed in a predetermined sequential manner. *Figure 13.11* portrays the distinction between these two addressing modes.

Another important characteristic of storage devices is the manner in which they are accessed during reading operations. The information in most devices is not altered during a read process, being classified as non-destructive readout (NDRO). The contents of some devices are altered or destroyed when reading, and are therefore classified as destructive readout (DRO) storage devices.

13.4.3 Semiconductor storage

The information in most RAM disappears when the power is turned off; therefore, it is characterised as 'volatile'. Storage devices which retain information after power down are 'non-volatile'. Dynamic RAM (DRAM) is RAM which is a DRO storage device. This is the fastest semiconductor memory device; however, it does require the additional overhead of refreshing memory due to its DRO nature. Static RAM (SRAM) is the class of semiconductor memory which does not

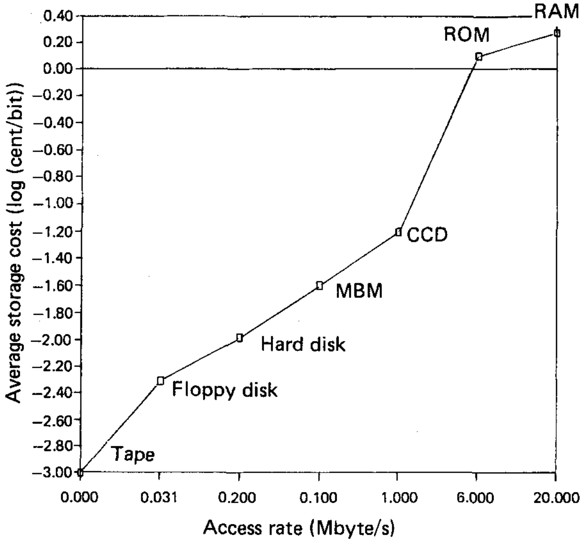


Figure 13.10 Storage category performance relationships

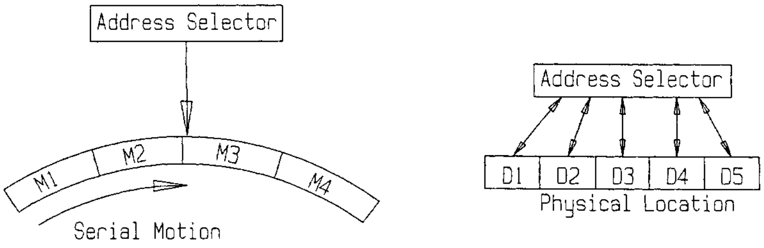


Figure 13.11 Conceptual address structures for SAM and RAM storage

require refreshing during reading operations (NDRO). This memory is still volatile, and has a slower access time than does the DRAM. It is frequently used as a temporary secondary storage device.

Recent technology has allowed for the development of non-volatile RAM (NVRAM); however, most RAM in this category is read-only memory (ROM). A ROM is a semiconductor device used for storing programs or other information as read-only information. ROMs function as input devices for dedicated functions. ROM is preprogrammed with a special device known as a 'programmable ROM' (PROM) burner. This process permanently alters the ROM. It should be noted that the terms PROM and ROM are interchangeable.

Two alterable versions of ROM are also available: the erasable ROM (EPROM), and the electrically erasable ROM (EEPROM). The EPROM is erased by exposure to ultraviolet light while the EEPROM is erased by a special control signal. All ROM devices are normally programmed off-line in a special programming unit. ROM technology has been an extremely cost-effective means of non-volatile storage of programs or dedicated instructions. This is available at a low cost per bit storage.

Two other semiconductor technologies have been used for storage. Magnetic bubble memories (MBM) provide non-volatile storage at a medium cost per bit, as shown in *Figure 13.10*. The bits in a MBM are stored as magnetic bubbles in cylindrical regions. These bubbles are moved in a serial manner to access the appropriate address for read/write operations. Charge coupled devices (CCD) operate in a similar fashion, charges being stored in an array of capacitors. The access rate from CCDs is faster than from MBM; however, CCD are volatile devices, and are therefore only suitable for temporary secondary storage.

13.4.4 Magnetic storage devices

Magnetic media have been used as traditional storage media for computers for some time. These devices include floppy disks, magnetic tape, magnetic drums and magnetic disks. Magnetic-tape units were the first magnetic storage devices used in computers, and still provide a cost-effective method of information storage. Most magnetic tapes have slower access times than magnetic disks and drums; therefore, most tape units are used as a tertiary back-up device.

The floppy disk is another popular cost-effective storage magnetic device. Data are storage in a serial manner in concentric tracks which are divided into sectors, as illustrated in *Figure 13.12*. The disk spins to provide serial access to the different sectors in the circumferential direction, while the head is positioned to provide access to discrete tracks in the radial direction. This addressing technique provides for the

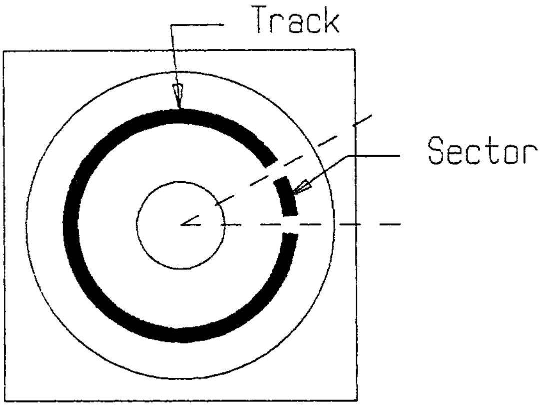


Figure 13.12 Floppy disk data address structure

addressing of data in a rectangular coordinate scheme, even though data are physically stored on a radial disk. Although the data are accessed in a serial manner, they are addressed directly and the floppy disk is characterised as direct access memory.

The predecessor of the floppy disk was the hard disk. Access is similar to that of the floppy disk, except the data are stored in much greater density and, therefore, less mechanical movement is needed to access data. Unlike the floppy disk, there can be several hard disks to a hard disk unit, or pack. Since there are multiple disks, there will be identical tracks in the same disk unit. The groups of identical tracks are termed 'cylinders' (see *Figure 13.13*).

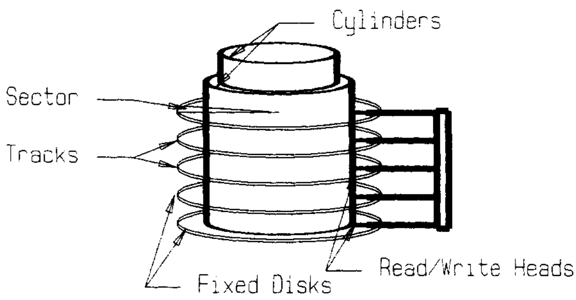


Figure 13.13 Moveable-head hard-disk cylinder with address structure

These disk packs are usually sealed, ensuring higher reliability; however, many of the larger systems provide for removable disk packs. Hard disks are not intended for frequent removal and are more expensive to use in terms of cost per bit than are floppy disks as previously indicated in *Figure 13.10.* This technology does provide a singificant increase in access speed and is the most popular mode of secondary memory.

The removable disk packs and many of the permanent packs use a movable head, although some units use a fixed-head system with multiple read units (see *Figure 13.14*). The fixed-head system provides an increase in reliability, speed, and storage capacity with a corresponding increase in cost per bit.

13.4.5 Hard copy output devices

Most hard-copy devices are classified as printers or plotters, where the data are transferred to paper in some fashion. Other methods for obtaining hard copy include videotape records of graphics images. Other photographic processes may be included for permanent hard-copy storage.

Graphics images are frequently stored as hard copy due to the tremendous amount of storage required to represent the image. These images may be stored during the development of a project; however, system managers will eventually be forced to erase these images to make room for more. This housekeeping of a system's storage facilities requires the back up of old information. Hard copy results may also be used for ease of transportability of the qualitative information.

The resolution obtained with the various hard-copy devices varies, and is a heavily advertised subject. One of the most economical devices is the dot-matrix printer. The dots are actually miniaturised solenoids patterned in a matrix formation. These solenoids actuate small strikers which impact an ink ribbon, leaving an ink dot on the paper. The resolution of the printer is determined by the number of dots in the print head and the size of the print head. The print head may be shifted slightly to fill the gaps between dots while printing. This mode is commonly referred to as 'letter quality'. The letter-quality mode normally requires two passes of the print head to print each line of text. Graphics images are bit mapped onto the page, requiring an extensive number of passes per inch of the print head. The printing of graphics images is time consuming due to the extensive movement of the print head.

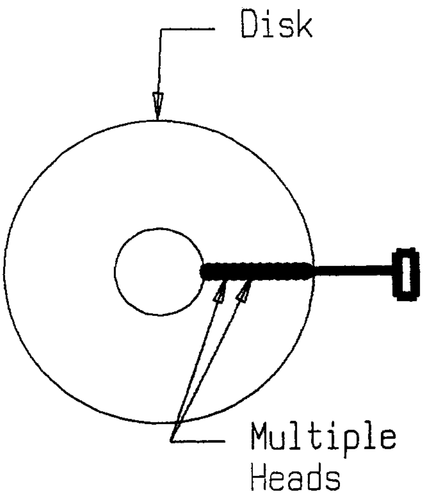


Figure 13.14 Fixed-head hard-disk cylinder

High-speed dot-matrix printers have a print head which does not move. These print heads are constructed of a series of dot-matrix heads in line, allowing the simultaneous printing of a line of text. The only mechanical movement required is the movement of the paper past the print head.

One of the most recent technologies implemented for hard-copy devices is the laser printer. Laser printers provide very high-quality output at moderate cost. The speed of these printers is normally higher than that of dot-matrix printers, and the operational cost is higher. These printers are frequently used for graphics output devices due to their high quality.

13.5 Human–computer interfaces

Many computer-controlled functions within a manufacturing environment operate without continuous interaction with a human operator. These systems use preprogrammed, self-opoerating functions which require initial programming input by a human. However, many other computer-controlled manufacturing functions use continuous human input to monitor and control various manufacturing operations. The interface hardware and software components between a computer and a human operator depend on several factors including:

(1) the manufacturing task to be monitored and/or controlled, including throughput timing and numbers of tasks to be monitored;
(2) the level and sophistication of the computer hardware and software;
(3) the number and type of tasks required of the human operator; and
(4) the degree of accuracy and quality control required of both the computer and the human operator.

The interface between a human operator and a computer requires communication hardware and software. These can be categorised as either (1) computer output devices which communicate the state of computer operation to the human operator, or (2) computer input devices which allow human input to control computer-driven manufacturing functions.

13.5.1 Computer output interface devices

13.5.1.1 Terminals, monitors and displays

The vast majority of peripheral devices which communicate the state of computer function to a human operator use visual output such as terminals, monitors and text and/or graphics displays. Many of these monitors are generic in nature and can be interconnected to various types and levels of computer. Other outputs are specialised in performance and can only be integrated into single-function computer systems.

Factors which should be considered when using visual displays include:

(1) whether the computer system is generic or specialised,
(2) the level of text information to be displayed,
(3) the level of graphics information to be displayed,
(4) the size of the display unit,
(5) the available space for a display unit,

(6) the level of ambient light and glare, and
(7) the distance between the display and the human operator.

Due to the developments in computer technology over the last decade, it is usually possible to match the computer display to the intended use and for an intended environment. Display systems can be purchased that are text specific, graphics intensive, and have touch and/or light-pen input capabilities. Large display systems are available for distant viewing, and glare screens are available for areas where the ambient light may otherwise adversely affect viewing of the screen.

13.5.2 Computer input interface devices

13.5.2.1 Keyboards

The keyboard represents the most widely used device for transmitting text information from a human operator to a computer. The design and operation of many keyboards have been standardised over many years of usage as data-entry devices. A keyboard may be a standard model which is commonly used with all levels and types of computer, or may be specialised for use with a specific computer or specialised equipment. The use of a standard keyboard as a data-entry device offers several advantages, including:

(1) flexible text input,
(2) user-defined function keys,
(3) independent numerical keypad,
(4) independent screen cursor control,
(5) simplicity of operation, and
(6) standardised hardware connection to the computer.

A specialised keyboard offers such advantages as predefined multifunction keys and user-specific layouts.

The most common keyboard layout uses the QWERTY system (see *Figure 13.15*). This system is named after the left-most six keys on the top row, and has become an international standard for keyboard layouts in English-speaking countries.[3] Variations of this system exist in France (as AWERTY) and in German-speaking countries (as QWERTZ). This layout is similar to that of typewriter keyboards with commonly used letter keys placed in close proximity. Approximately 10 key postions out of a total of 100 available positions (a combination of 50 lower and upper case) can be user specific for specialised applications (see *Figure 13.16*).

A separate numerical keypad is located to one side, as is a separate screen-cursor-control keypad. The numerical keypad is arranged in a 3 number by 3 number array with the zero at the top or the bottom of the array (depending on whether the array is arranged in ascending or descending numerical order). In addition, a full number set is situated as lower-case keys on the top row of the main key area. The cursor-control layout usually appears in a T-shaped or cross-like arrangement. The latter configuration usually contains a 'home' key which, when pressed, sends the cursor to top left-hand corner of the computer display.[4]

A set of function keys may be placed at either the top or on the left-hand side of the keyboard. The use of top-mounted function keys is more appropriate for keyboards placed to the left of the user because this arrangement minimises the reach distance of the human operator. These function keys may be user defined or may have predetermined functions, depending on the specific application software in operation.

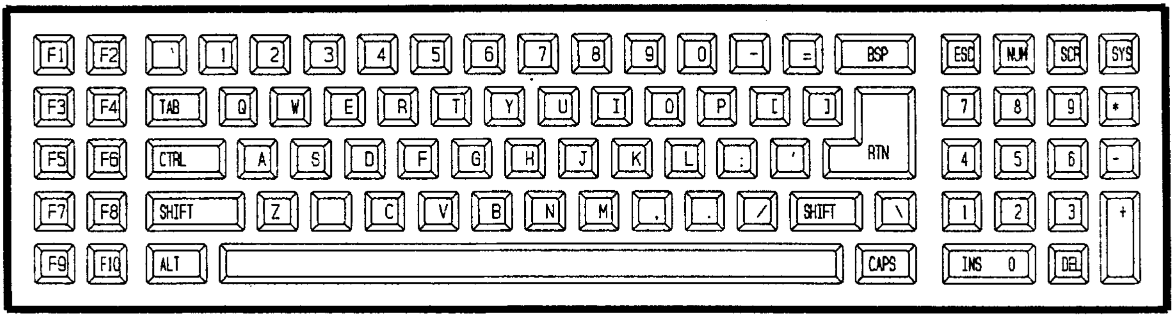


Figure 13.15 QWERTY keyboard

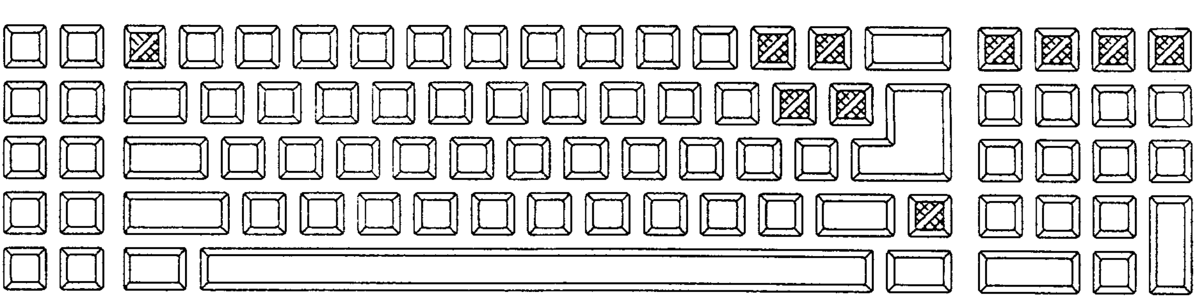


Figure 13.16 QWERTY key arrangement with user-specific keys

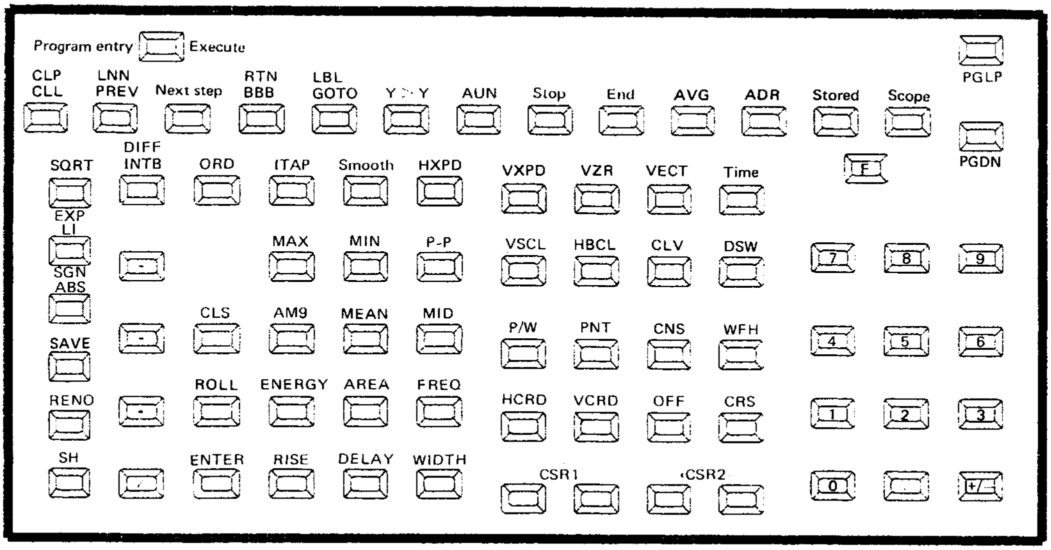


Figure 13.17 Specialised keyboard

The DVORAK keyboard system is a common variation on the QWERTY system, with a few letter and punctuation keys moved to different keyboard locations in order to allow both hands to use the keyboard simultaneously without excessive awkwardness.[4] However, for computer applications where the operator will be using only one hand to input data, the QWERTY system is more advantageous. This system overloads the left hand, and thus a QWERTY keyboard is more appropriately placed to the left of the operator.

Standard keyboard layouts can also vary in terms of keyboard dimensions, key sizes, and tactile sensation. A typical key size is 13 mm × 13 mm with 19 mm spacing (centre to centre) between keys. The key-activation force varies from 0.25 to 1.5 N with key displacement varying from 1 mm to 6 mm. The selection of a keyboard interface should incorporate the intended application, including anticipated key force and whether a bare or gloved hand will be used to input data.[5]

An example of a specialised keyboard is shown in *Figure 13.17.* Specialised keyboards are usually smaller than standard keyboards and are less flexible. There are fewer keys and the function keys are predefined for the intended application. Such keyboards are routinely used as an interface to computerised electronic testing equipment, as well as for robotic controllers.

Considerations for keyboard selection and usage include the placement of function keys, numerical keys and cursor-control keys as well as overall keyboard dimensions, key size and required key force.[6,7] The keyboard should be placed near the attached computer in order to minimise interference of the keyboard connecting wires with on-going operations. This, in turn, should affect the selection of function-key placement design. The dominant hand of the user will also affect the optimum placement of the keyboard.

Keyboard usage is most advantageous for input of text or specialised functions via function keys. Thus, its use as a manual computer interface device is intended for situations where variability and flexibility of alphanumerical input is most important, or where specific, single-key functions can be elicited. It is not advantageous in situations where the human operator does not have at least one hand free or where the computer function is menu driven or graphics intensive.

13.5.2.2 Mice and joysticks

A mouse (*Figure 13.18*) is a hand-operated input device which is attached by a wire to a computer through a serial communications port. Typically, it is a small, rectangular plastic box which is placed under the palm of the user's dominant hand and is slid along a flat surface to control the cursor position on a computer screen. A mouse contains one to three buttons on the top which can be clicked by the index or middle finger to activate functions such as selecting computer screen menus, drawing objects on screen, or specifying selected menu options. A mouse may be either mechanical or optical in its operation. Both types operate by converting sensor information from mechanical wheels or optical sensors into x–y coordinates of cursor position on the computer monitor/screen.[8–10]

A mouse is a rapid cursor-control device. Its use in the operation of menu-driven software is far superior to that of the

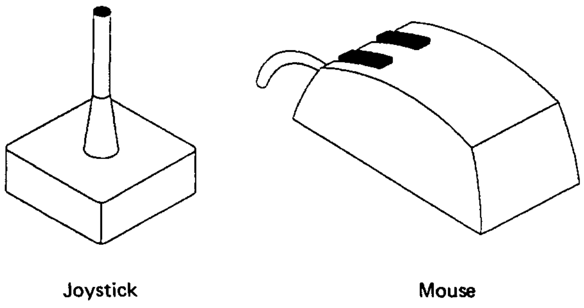


Figure 13.18 Mouse and joystick

cursor keypad controls on a standard computer keyboard.[11] However, a mouse is exclusively intended as a cursor-positioning device and as a limited input-selection device. For those applications which require more flexible inputs, a mouse cannot be used as the sole data-input system. In addition, the use of a mouse requires additional table space over and above that of a keyboard (usually a 20 cm × 20 cm square). The advantages of a mouse include its rapid control of cursor position, the ability of the user to look at the screen without the need to look down at the mouse during operation, and its use as a free-hand-drawing device. However, usage of a mouse for accurate drawing purposes is limited as the mouse is somewhat awkward as a drawing tool and its resolution is limited. It is an excellent input device for applications which are primarily menu driven, including CAD systems. *Table 13.2* lists various mouse manufacturers and the features and relevant supported software systems of commonly used mice.

A joystick (*Figure 13.18*) is a lever which is mounted vertically in a fixed base. It operates by tilting the lever forward, backward, left or right in order to control the cursor position on a computer screen in much the same fashion as the control elicited by a mouse. A button on the top of the lever is used as a single input-selection device, similar to that of a mouse button. A joystick is connected by wire to a serial communications port of a computer similar to that of a mouse. A joystick is advantageous where there is limited table space available, in that one can control the cursor position from a single spot, whereas a mouse requires hand movement over a small area.[12] However, unlike a mouse, a joystick cannot be used as a drawing device. It is primarily used for cursor position and menu selection. A typical mouse and joystick are shown in *Figure 13.18*.

Both mice and joysticks are low-cost devices with limited usefulness. Their resolution is limited. Thus, the number of menu selections and the size of each menu item displayed on screen may be a limiting factor in the use of either of these pointing devices.[13,14] The use of a mouse for long periods appears to be less tiring than the use of a joystick, which eventually affects the wrist muscles of the user.

13.5.2.3 Touch screens

A touch screen provides a direct method of cursor-position input to a computer monitor or screen in response to a touch of a finger on the screen at the intended location. It is primarily used as a menu-selection system where the menu is placed in a one- or two-dimensional linear array, as shown in *Figure 13.19*. This differs from the typical use of a mouse as a menu-selection device in that mouse-driven systems usually use pull-down menus from a highlighted band of selections at the top of a computer screen. Touch screens can operate with physical screen overlays, or by conductive, optical, ultrasonic or physically sensitive grids that are incorporated into the computer screen, as shown in *Figure 13.20*. Because a touch screen is a specialised computer monitor which is quite different from a standard computer monitor, its usage is

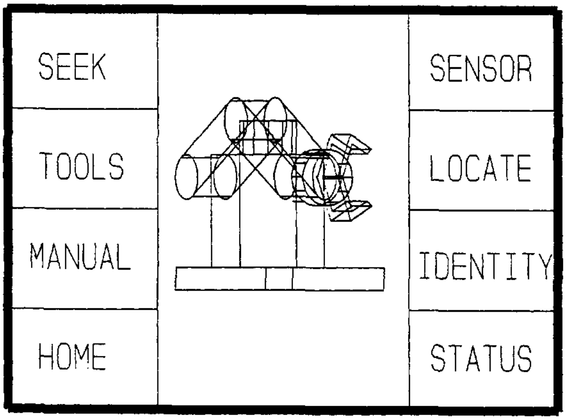


Figure 13.19 Menu-driven computer screen

Table 13.2 Mouse features

System	*Sensor*	*Interface*	*Buttons*	*CAD Support*	*Windows support*
Logitech Mouse, Logitech Inc.	Optomechanical	Serial, bus	3	Yes	Yes
Microsoft Mouse, Microsoft Corp.	Mechanical	Serial, bus	2	Yes	Yes
PC Mouse, Mouse Systems Corp.	Optical	Serial, bus	3	Yes	Yes
IBM PS/2 Mouse, IBM Corp.	Mechanical	Special-ised	2	Yes	Yes
SummaMouse, SummaGraphics Corp.	Optical	Serial	3	Yes	Yes
IMSI Mouse, IMSI Inc.	Optical, mechanical	Serial	3	Yes	Yes
American Mouse, American Computer and Peripheral Inc.	Mechanical	Serial, bus	3	Yes	Yes
MultiMouse, American West Engineering Corp.	Optomechanical	Serial keyboard	6	Yes	Yes
Manager Mouse, Torrington Co.	Mechanical	Serial	3	Yes	Yes

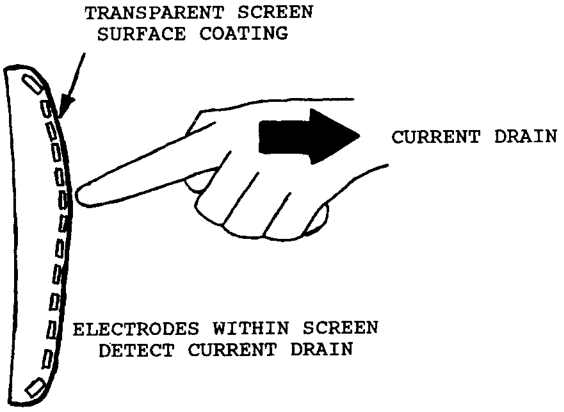


Figure 13.20 Touch-screen operation

intended for specialised applications where a standard monitor is not required. Thus, the use of alternative input devices (which require a standard computer monitor) in concert with a touch screen is usually precluded. There are a few touch-screen systems which can be installed on standard computer monitors, as opposed to purchasing a dedicated computer/touch-screen system. However, their performance in terms of resolution and speed of operation is inferior to that of dedicated systems.[9]

A major advantage of the touch screen is that the hand–eye coordination is maximal because the input and output devices are one and the same, namely the computer monitor screen. Since the number of menu selections and functions is limited, there are usually far less errors in the use of a touch screen as an input device than occur with other input devices such as a keyboard or mouse. The software that is incorporated for use with touch screens is often designed to minimise user training. Thus, the use of touch screens is often much simpler than that of other input devices in that the user is often 'led through' a series of menus leading towards the selection of a particular computer-controlled function.[15]

However, there are also several disadvantages to the use of touch screens. Arm fatigue is common after even moderate use, since the operator must usually lift an arm in order to reach the computer screen. A touch screen is limited as a computer input device to menu-driven applications. This is typical of some robotic functions and many electronic testing applications. Data entry is slowed by arm movement and the screen is temporarily blocked from view as a menu selection is made. The selection of small items on a screen may also be problematic. It is possible inadvertently to select the wrong menu item, especially after prolonged use. Due to constant contact with fingers, touch screens require frequent cleaning in order to remain effective.

Touch screens are not available for all computer systems. They are typically used with selected microprocessor-based systems and have seen only limited use as a computer input device, primarily as an adjunct to computerised electronic circuit-testing systems. Some keyboards may be purchased with small, specialised, user-definable touch pads incorporated in the lower right-hand-side of the keyboard, near the cursor-control keypad.

13.5.2.4 Light pens

A light pen is essentially a pencil-sized electronic version of one's finger which is used to point to a menu selection on a light-sensitive computer screen. Thus, it is quite similar in function to a touch screen. The use of a light pen requires a light-sensitive CRT display which is quite different from a standard computer monitor.[9] Thus, as with a touch screen, the use of a light pen often precludes the use of alternative input devices. A typical light pen is shown in *Figure 13.21*.

As with a touch screen, a light pen is most often used as a pointer to select items from a menu array displayed on screen. One merely points at a particular menu selection and enables the pen by a push button located at its tip. However, a light pen can also be used as a drawing device by dragging the pen along the screen while it is enabled. Since there is not direct contact of fingers with the computer screen, the screen does

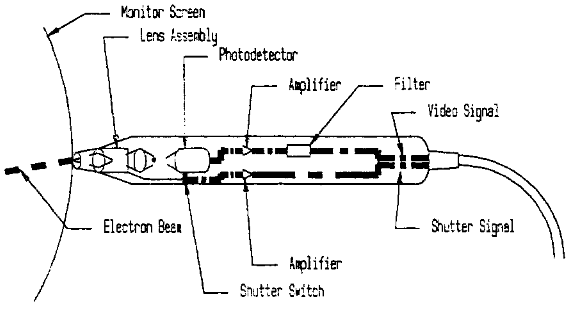


Figure 13.21 Light-pen operation

Table 13.3 Light pens

System	*Graphics adapter*	*CAD support*	*Windows support*
L-PC Lite Pen, The Light-Pen Co.	CGA, EGA, mono.	No	Yes
FT-156 Light Pen, FTG Data Systems Inc.	CGA, EGA	Yes	Yes
Fast-point Light Pen, HEI Inc.	CGA	Yes	Yes
Warp Speed Light Pen	CGA, EGA, mono.	Yes	Yes

CGA, colour graphics adapter; EGA, enhanced graphics adapter; mono., monochrome.

not need frequent cleaning as does the touch screen. *Table 13.3* lists the major manufacturers and the features of commonly used light-pen systems that are additions to otherwise standard computer systems. It is also possible to purchase a dedicated computer system with a built-in light-pen system that is designed for specialised applications such as electronic-circuit testing and CAD systems.[16]

The disadvantages of the light pen are similar to those of the touch screen. Arm fatigue is common. The screen is blocked during the menu-selection process. Resolution is limited and incorrect selections are quite possible. In addition, light pens and light-sensitive computer screens are not common and are not available for all computers. As with the touch screen, light-pen systems are marketed for selected microprocessor-based systems. Typical applications are menu-driven functions such as those found with electronic-circuit-testing systems.

13.5.2.5 Graphics tablets and digitisers

A graphics tablet is a flat panel which is connected by a wire to a serial communications port in a computer. These are typically available only for microprocessor-based systems. A pencil-sized stylus is attached by a wire to the tablet and is dragged along the tablet surface. The tablet mimics the computer screen in that the path of the stylus is depicted directly on the screen. Thus, a graphics tablet is primarily used as a graphics-input device. However, unlike a mouse or light pen, the use of a graphics tablet allows drawing to be accomplished horizontally in a manner similar to one's natural drawing mode on a flat horizontal surface.

A graphics tablet is useful for drafting and drawing with the input appearing directly on the screen. The resulting graphic display can be stored in the computer memory for later retrieval and modification. One can trace an existing drawing by placing it onto the tablet and dragging the stylus over the drawing.[17] A drawback in its usage is that the tablet and the resulting display on the computer screen are physically separated. Thus, there is limited hand–eye coordination between what is being drawn and the resulting image on the screen. In addition, a graphics tablet requires considerable table space (typically a 50 cm × 50 cm square) and has limited resolution.[18] Data entry is slower than with many other input devices. *Table 13.4* lists the manufacturers and the features of commonly used graphics tablets.

An alternative method of direct graphic data entry is a graphic digitiser. This device scans an existing paper copy and digitises the image for storage in computer memory. Such a system offers greater resolution than a graphics tablet, albeit at considerable expense. However, a scanning digitiser allows direct entry of circuit diagrams and similar detailed graphic information into computer memory. Both the graphics tablet and the scanning digitiser are primarily graphic-image generators and their application in a computerised manufacturing environment is limited.

13.5.2.6 Eye position and gaze input

The use of eye position as an input to a computer is primarily intended for applications where both hands are occupied with other tasks and are not available for data entry by keyboard, mouse or other input device. Such a system requires the use of specialised glasses which are connected by wires to the computer through an electronic interface. The computer connection is similar to that of a mouse, which is accomplished through a serial communications port. An eye position input system is also similar to a mouse in that it acts as a cursor positioning device. However, as it is intended as a 'hands-free' system, there is no analogue to the mouse button. Instead, switching and selection is accomplished by fixing one's gaze on a single spot on the computer screen for a fixed time period, which is selectable by the operator.

An eye position input system operates via detection of infra-red light which is shone into and reflected from the eye. An alternative system tracks eye position through the us of light emitting diodes (LEDs). Such systems require calibration of the operator's gaze with computer screen cursor position. These systems are heavily dependent on head motion. They are primarily intended for use in the selection of menu items on screen, in a similar fashion to that of the touch screen, light pen, or mouse. As the time required to fix one's gaze in selecting a menu item is programmable, the input response time is variable. However, currently employed systems can produce menu selections in less than 0.2 s, which is much faster than the response time for touch screens or mice, which incorporate a relatively slower hand motion in the selection process.[19]

The use of eye position systems requires specialised equipment which is relatively expensive compared with other computer input devices. A major disadvantage is the need for specialised glasses which may interfere with other operator tasks. However, the use of eye position input devices to control computer functions allows 'hands-free' operation, unlike other input devices. The technology of these systems is still relatively young compared with that of other input

Table 13.4 Graphics tablets

System	*Size* (cm × cm)	*Programmable*	*CAD support*	*Windows support*
IS/ONE Series Tablet, Kurta Corp.	41 × 38	Yes	Yes	Yes
VersaWriter, Versa Computing Inc.	30 × 34	No	No	No
SummaSketch 1201, SummaGraphics Corp.	41 × 41	No	Yes	Yes
KoalaPad + Touch Tablet, Koala Technologies Inc.	15 × 20	No	No	No
Digi-Pad Type 5A, GTCO Corp.	40 × 40	Yes	Yes	Yes
Pencept Penpad 320 Pencept Inc.	38 × 38	Yes	Yes	Yes

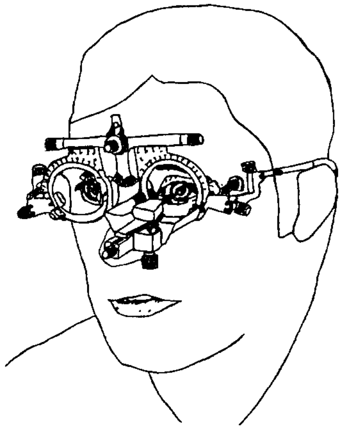


Figure 13.22 Specialised glasses for use with an eye-controlled computer system

devices. However, the use of eye tracking systems for such applications as aircraft controls has advanced the technology in this area to such a degree that it may become a premier 'hands-free' computer control system in the near future.[20] A typical eye position control system is shown in *Figure 13.22.*

13.5.2.7 Voice input

The use of human speech as an input to control computer function represents the optimum 'hands-free' methodology presently available. The use of speech as a controlling input allows the operator to use his/her hands for other functions and to gaze at the computer monitor while issuing commands. Speech-recognition systems are available from many commercial sources and can be interfaced to various types of computers, including mainframe, mini- and micro-computers. However, most of these systems are designed for application with microcomputer-based systems, where a solid-state expansion card fits into an existing personal computer (PC). Other systems operate as stand-alone systems with a built-in microprocessor which is connected to an existing computer through a serial communications port. The expansion card versions of speech-recognition systems are slower, but are also far less expensive than stand-alone models.

The voice signal is monitored by a microphone which is incorporated into either a light-weight headset worn by the operator or placed on a table-mounted free-standing base. Frequency modulated (FM) telemetry may be employed in conjunction with the headset and a battery powered FM amplifier may be used to allow the operator to move about without being physically tethered to the computer. Thus, robotic functions, assembly-line operations, materials (inventory) handling and hazardous materials handling may all be controlled by voice through computer controls.[21]

The voice signal is transduced by the microphone and preprocessed by an analogue filter/amplifier network in order to eliminate spurious ambient noise. This analogue signal is then digitised via an analogue-to-digital converter within the processing computer. The digital signal is then divided into its frequency components by either a digital filter network or through the use of a fast Fourier transform (FFT) or other time domain-to-frequency domain calculation. Each uttered word or brief phrase is thus transformed into a stored pattern of amplitude vs. frequency. Typically, the processed speech is limited to the telephone bandwidth (300–3000 Hz).[22–25]

A working vocabulary must be created which is relevant to the intended application and whose words are phonetically different so that their stored patterns are dissimilar. In general, short words with similar sounds, such as 'go' and 'no', are not dissimilar enough to be readily distinguishable by the speech recognition system, and should therefore not be included in a working vocabulary as a group. The frequency components of the entire working vocabulary are stored as a series of templates during a brief vocabulary training period. For many speech recognition systems, the training period may be as brief as 1 min, depending on the vocabulary size. Due to the variability between individual voices, all operators who would use a voice control system must train and record their respective vocabulary templates. Once this has been done, an operator simply keys in his identity so that the system calls up the proper template. During control operation, each spoken word is compared with the stored template to determine the proper word match. Once this has been accomplished, an appropriate process control function may be initiated by the computer. The processing speed of such systems is rapid, such that the human operator can initiate 'real time' control of computer driven manufacturing functions.

Voice processing systems are evaluated on the basis of:

(1) overall word recognition performance,
(2) training requirements,
(3) vocabulary capabilities,
(4) noise immunity, and
(5) computer interface requirements.

A typical voice processing system contains a microphone, a noise cancelling system, an analogue-to-digital converter, a time-to-frequency converter, a frequency template development system, a training module, a performance analayis module, and perhaps a telemetry system for remote operator utility (see *Figure 13.23*). *Table 13.5* lists the concerns which must be addressed for each component of the voice recognition system.

In particular, microphone directionality, noise immunity, and processed word error reduction are of considerable concern to the operation of a voice recognition system as an integral part of a computerised control system for a manufacturing environment. When using a table-based microphone, the speaker must take great care to speak towards the microphone in order to reduce the errors associated with speech recognition. Even when using headset-mounted microphones, it is important for the operator to speak towards the microphone, because many such systems employ swing away microphones. In addition, the tone of the operator's voice should remain relatively constant in order to ensure proper accuracy in word recognition. Thus, operations which might severely affect the tone of the human voice would be poor candidates for speech controlled processes.[26,27]

Noise cancellation in an industrial setting is a significant concern in the use of speech processors. Industrial noise may incorporate broadband noise resulting from blowers and fans, discrete frequency noise resulting from various periodic sources, and aperiodic impulse noise resulting from stamping and pressing machines and related sources. Centre clipping of the voice signal as well as other data-processing methodologies have significantly reduced the effects of broadband noise on processed speech. Periodic noise has been successfully treated with predictive filtering techniques. However, impulse noise of sufficient amplitude to mask the voice signal remains a serious problem. Noise cancellation may be accomplished through an analogue circuit placed between the microphone and computer interface, or by means of digital processing within the computer itself.[28,29]

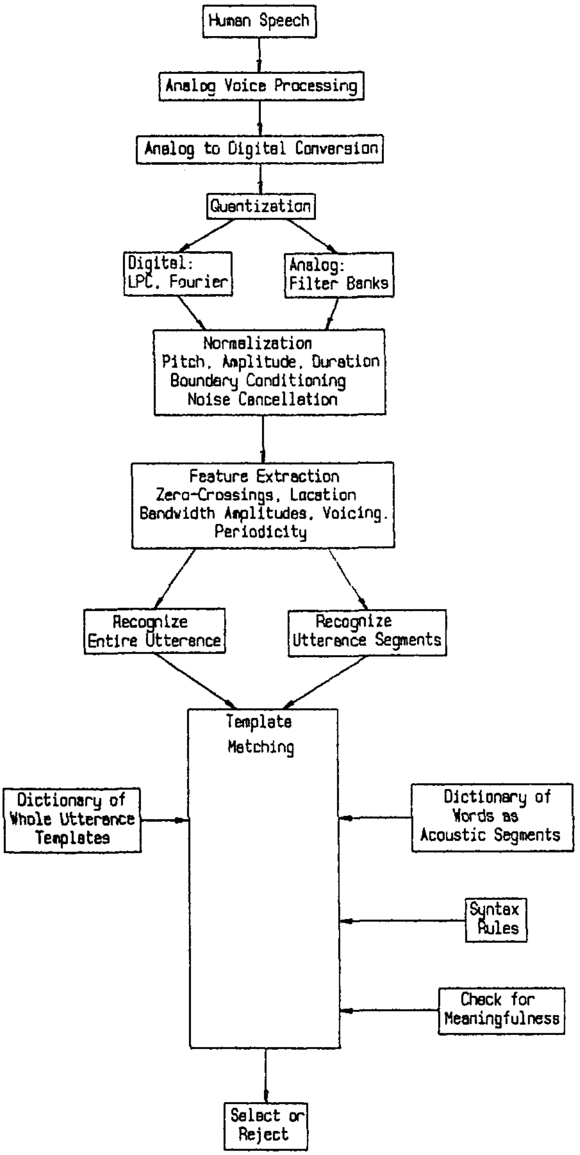


Figure 13.23 Voice processing system. LPC, linear predictive coding

Errors resulting from incorrect voice processing may be due to either omission or commission. Omission errors are those in which the spoken word is not recognised and no computer control command is elicited. Commission errors are those in which the wrong word in the template is matched and an incorrect computer control command is elicited. Correct noise cancellation techniques can significantly reduce the incidence of errors during speech processing. Virtually error-free operation has been demonstrated for broadband noise up to 110 dB.

Voice processing as an input to computer-controlled manufacturing processes is a reliable method of 'hands'-free' input and can serve as a useful adjunct to other input devices such as a keyboard or mouse. Voice based systems can serve as a keyboard emulator, thus producing any input of a standard keyboard entry, or as a mouse emulator to provide menu selection operation. Code words or short phrases can be used as an alternative to user-defined function keys on a keyboard. *Table 13.6* lists the most commonly used computer based speech recognition systems and their respective features.

Speaker independence, impulse noise cancellation, and limited vocabulary size are areas where future research is targeted in order to expand voice-controlled applications in a manufacturing environment.[30–32]

Table 13.5 Areas of concern for speech recognition systems

System component	*Areas of concern*
Voice processing board	Speaker dependency Speed of processing Bandwidth Noise cancellation method Word recognition method Training requirements Accuracy Vocabulary
Microphone	Type of element Directional sensitivity Amplitude sensitivity Noise cancellation
Computer interface	Analogue-to-digital conversion rate Analogue-to-digital conversion
Accuracy	Power requirements Connection slot or bus Bandwidth
Telemetry	Bandwidth Amplitude sensitivity Directional sensitivity Accuracy

13.6 Networks

The term 'computer network' is used to describe situations in which: (1) physically remote terminals are connected to a central processing computer, (2) many remote small computers are interconnected with the aid of a central processor or storage unit, or (3) independent large computer systems are linked in order to share hardware peripherals or software systems.

A glossary of network system terms is given in *Table 13.7*.

13.6.1 Basic communications classification

The flow of information is critical in modern manufacturing environments. Many processes are now controlled or synchronised by a host computer or by several computers in an integrated system. Speed and reliability are critical to these information networks.

The flow of information between two devices may be categorised as one-directional (simplex), alternating informa-

Table 13.6 Voice recognition systems

System	*Unit placement*	*Uses own microprocessor*	*Training passes required*	*Vocabulary size*
Voice Scribe 1000, Dragon Systems Inc.	Internal	No	5	1000
Voice Card VPC2150, Votan Corp.	Internal	Yes	1	64
Kurzweil Voicesystem, Kurzweil Applied Intelligence Corp.	External	Yes	3	1000
IBM Voice Communications Option, IBM Corp.	Internal	Yes	5	160
TI-Speech System, Texas Instruments Inc.	Internal	Yes	3	50
KB5152V, Key Tronic Corp.	External	Yes	3	160
Pronounce, Microphonics Technology Corp.	Internal	No	1	128
Voice Developer System, Verbex Voice Industries Corp.	External	Yes	4	100
VocaLink SRB-LC Interstate Voice Products Inc.	Internal	No	3	400
VoiceKey, Roar Technology Inc.	Internal	No	2	512
VoiceCommand, Interpath Corp.	Internal	No	4	200
IntroVoice VI, The Voice Connection	Internal	No	3	400
Coretechs VET 3 Speech Terminal, Scott Instruments Corp.	External	Yes	3	200

Table 13.7 Glossary of network system terms

Term	*Description*
Access protocol	The traffic rules that LAN workstations abide by to avoid data collisions when sending signals over shared network media. Common examples are carrier sense multiple access (CSMA) and token passing
APPC	Advanced program-to-program communications: this IBM protocol is analogous to the OSI session layer, since it sets up the necessary conditions that enable application programs to send data to each other through the network
Application layer	The highest level of the OSI model. It describes the way that application programs interact with the network operating system
ARCnet	Attached resources computing: a networking architecture using a token-passing bus architecture, usually on coaxial cable
Bridge	A type of product that links different local area networks, enabling users on one network to use all the resources available on the other

Table 13.7 Cont'd

Term	*Description*
Buffered repeater	A device that amplifies and regenerates signals so that they can travel further along a cable. This type of repeater also controls the flow of messages to prevent collisions
Bus	A 'broadcast' arrangement in which all network stations receive the same message through the cable at the same time
Cache	An amount of RAM set aside to hold data that network stations are expected to access again. The second access, which comes from RAM, is very fast
Coaxial cable	A type of network medium. Coaxial cable contains a copper inner conductor surrounded by plastic insulation and then a woven copper or foil shield
CSMA	Carrier sense multiple access: a media-sharing scheme in which stations listen in to what is happening on the network media; if the cable is not in use, the station is permitted to transmit its message

Table 13.7 Cont'd

Term	*Description*
Data link layer	The second layer of the OSI model. Protocols functioning in this layer manage the flow of data leaving a network device and work with the receiving station to ensure that the data arrive safely
Ethernet	A network cable and access protocol scheme originally developed by Xerox, now marketed mainly by DEC and 3Com
Fibre optics	A data transmission method that uses light pulses sent over glass cables
File server	A type of server that holds files in private and shared subdirectories for LAN users
Frame	A packet on a token-ring network
Gateway	A portal though which networked stations can access resources on mainframe and minicomputer hosts as well as on nationwide data networks
ISO	International Standards Organization (ISO), Paris, which developed the open system interconnection (OSI) model
Locking	A method of protecting shared data. When an application program opens a file, file locking either prevents simultaneous access by a second program or limits such access to 'read only'
MAC	Media access control
MAP	Manufacturing automation protocol: a token-passing bus LAN designed by General Motors
Media	The cabling or wiring used to carry network signals. Typical examples are coaxial, fibre-optic, and twisted-pair wire
NetBIOS	Network basic input/output system: a layer of software originally developed by IBM and Sytek to link a network operating system with specific hardware. It can also open communications between workstations on a network at the session layer
Network layer	The third level of the OSI model, containing the logic and rules that determine the path to be taken by data flowing through a network
NFS	Network file system: one of many distributed file system protocols that allow computers on a network to use the files and peripherals of another networked computer as if they were local
OSI	Open system interconnection: a model developed by the ISO describing the network communications process
Packet	A block of data sent over the network transmitting the identities of the sending and receiving stations, error-control information, and a message
Peer-to-peer resource sharing	A software architecture that lets any station contribute resources to the network while still running local application programs
Physical layer	The first layer of the OSI model. It consists of network wiring and cable
Presentation layer	The sixth layer of the OSI model, which formats data for screen presentation and translates incompatible file formats

Table 13.7 Cont'd

Term	*Description*
Print server	A computer on the network that makes one or more attached printers available to other users. The server usually requires a hard disk to spool the print jobs while they wait in a queue for the printer
Protocol	A specification that describes the rules and procedures that products should follow to perform activities on a network, such as transmitting data. Protocols allow products from different vendors to communicate on the same network
Redirector	A software module loaded into every network workstation. It captures application programs' requests for file- and equipment-sharing services and routes them through the network
Repeater	A device that amplifies and regenerates signals so they can travel farther on a cable
RFS	Remote file service: one of the many distributed file system network protocols that allow one computer to use the files and peripherals of another as if they were local. Developed by AT&T and adopted by other vendors as a part of UNIX V
Router	A machine in a large network that reads the destination of a message and selects the best route
SAA	Systems application architecture: a set of specifications written by IBM describing how users, application programs, and communications programs interface
Server	(a) A computer with large power supply and cabinet. (b) Any computer on a network that makes file, print, or communications services available to other network stations
Session layer	The fifth layer of the OSI model, which sets up the conditions whereby individual nodes on the network can communicate or send data to each other. The functions of this layer are used for many purposes, including network gateway communications
SMB	Server message block: a distributed file system network protocol that allows one computer to use the files and peripherals of another as if they were local
SNA	System network architecture: IBM's model of a communications system
StarLAN	A networking system developed by AT&T that uses CSMA protocols on twisted-pair telephone wire
Token passing	An access protocol in which a special message (token) circulates among the network nodes giving them permission to transmit
Token ring	Refers to the wire and the access protocol scheme whereby stations relay packets around in a logical ring configuration
Topology	The map of the network. The physical topology describes how the wires or cables are laid out, and the logical or electrical topology describes how the messages flow

Table 13.7 Cont'd

Term	*Description*
Transport layer	The fourth layer of the OSI model. Software in this layer checks the integrity of and formats the data carried by the physical layer, managed by the data layer, and routed by the network layer
Twisted-pair wiring	Cable comprising two wires twisted together at 6 turns per inch to provide electrical self-shielding
X-25	An international standard describing how computers can access packet-switched networks. Typical X-25 networks include Tymmet, Telenet, and MCI Mail
XNS	Xerox network services: a multilayer protocol system developed by Xerox and adopted, at least in part, by Novell and other vendors. XNS is one of the many distributed file system protocols that allow network stations to use other computers' files and peripherals as if they were local

LAN, local area network.

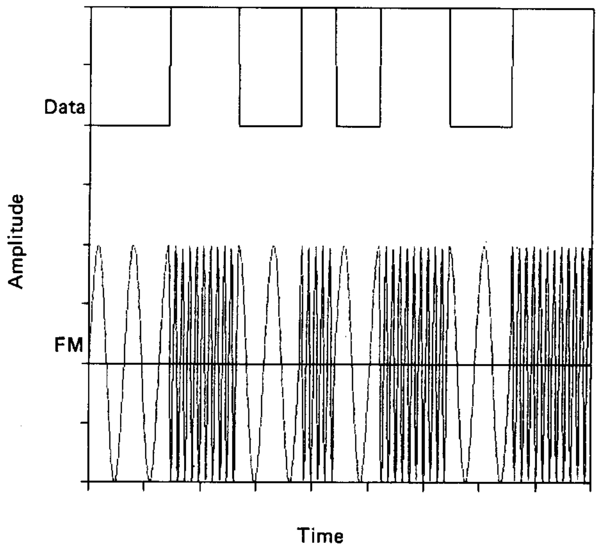


Figure 13.24 Frequency modulation

tion flow (half duplex), or both directions simultaneously (full duplex). A permanent data path between two discrete devices is dedicated, whereas a path with multiple devices is shared. The ideal system would have dedicated links between all devices; however, this is not normally physically or economically feasible. Therefore, most systems comprise a network of shared data links.

This flow of data may be modulated and demodulated (MODEM) with a carrier frequency. The carrier wave is altered (modulated) in accordance with the value of the input data. The frequency, phase, or amplitude are the most common parameters which are altered. *Figure 13.24* illustrates the frequency modulation (FM) technique. This modulated signal is demodulated to extract the data at the receiving point. This method extends the transmission distance of many signals. Frequency modulation techniques usually enhance the noise rejection characteristics of the system as well.

Data transmitted over phone lines normally have a carrier frequency of 1000 Hz. MODEM systems with one frequency are termed 'baseband', whereas systems with multiple carrier frequencies are classified as 'broadband'. A single data link between several transmission devices can transmit several different logical transmissions simultaneously at different frequencies. This data link can therefore provide for multiple logical paths over the same physical link using MODEM transmission techniques.

The flow of data in any system can be categorised as synchronous or asynchronous. The information flow inside a computer system is passed from one point to another at discrete instants controlled by a system clock. This method synchronises (synchronous) the data flow within the computer. When connecting several on-going processing units, an asynchronous link is normally used. This method permits the flow of data without the need to synchronise the transmitting and receiving devices. The synchronous system must interrupt the transmitting and receiving points to synchronise them and transfer the data. An asynchronous system transmits a start bit and stop bit to delineate the beginning and end of data words, respectively.

13.6.2 Transmission media

There are several different physical means for the transmission of data. *Table 13.8* lists some of the more common media with their respective data bandwidths. The data bandwidth of a system is defined as the difference between the minimum and maximum data frequency transmission rates. Fibre optic technology has been used extensively in recent years in industrial settings due to its immunity to electromagnetic interference. A fibre optic line cannot be 'tapped' easily, thus increasing data security while decreasing the ease of splicing for repair or expansion.

The baseband coaxial or twisted pair are still used extensively due to their relative cheapness and simplicity. These systems are much easier to repair, and are easier to modify for additional installations. However, this ease of repair and modification reduces the system security.

The implementation of either coaxial or twisted pair can be connected as an unbalanced transmission link, as shown in

Table 13.8 Data transmission medium characteristics

	Twisted pair	*Baseband coaxial*	*Broadband coaxial*	*Fibre optic*
Baud rate (Mbyte/s)	0.1	10	400	1700
Cost	Lowest	Low	Higher	High
Noise immunity	Poor	Intermediate	Good	Immune
Security	Lowest	Fair	Improved	Very high
Maintenance cost	Lowest	Low	Higher	High
Equipment complexity	Lowest	Low	High	Intermediate
Maintenance skill	Lowest	Low	High	High
Other	High transmission loss		Conduit somtimes required. MODEMs required	Small size. Point–point links. Difficult to repair

Figure 13.25 Unbalanced transmission link

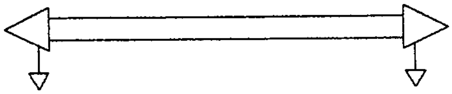


Figure 13.26 Balanced transmission link

Figure 13.25. The difference in ground potential (ground loop) between the transmitter and receiver may introduce noise into the system. The balanced link is depicted in *Figure 13.26*. The differential amplifier configuration will eliminate most ground loops as well as reduce electromagnetic noise in the signal.

The theoretical analysis of communications networks is closely related to systems theory analysis of manufacturing systems. Data networks transfer information at much greater rates than typical manufacturing systems, increasing the possibility for errors. Deliberate redundancy and inefficiency may be required in a data transmission system to ensure reliable transmission.

13.6.3 Network topologies

The topology of a network is functionally identical to the architecture of a computer. It represents the physical map of the data links in the network. A few of the common network topologies such as the star, distributed or ring configurations are illustrated in *Figure 13.27*. The star version of a network requires all communications to travel through a central element. A distributed version allows for alternative communication paths between elements. The ring version of a network requires communication through a circle of elements, in effect through all elements in the path between the transmitting and receiving elements.

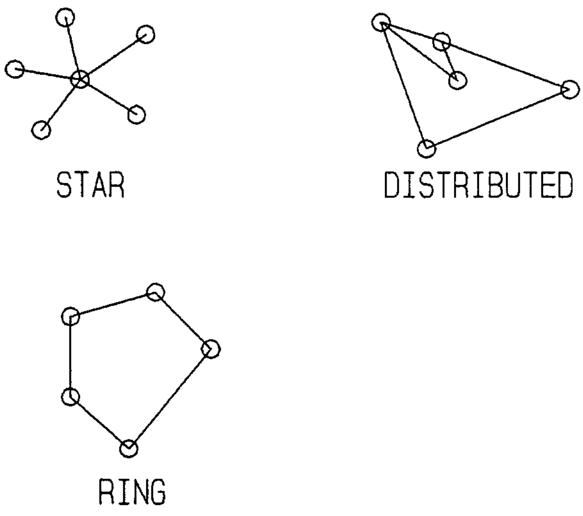


Figure 13.27 Typical network configurations

Networks are composed of *nodes* (the elements described above), with *circuits*, *channels*, or *links* connecting them. A node could vary from a small piece of computer hardware up to a large computer within the network. The channels could consist of hardwire connections, microwave link ups, satellite links, radio links or cable television links. However, the method of linking is usually transparent to the user. Messages between two modes are usually split into small *packets* of finite length. Packets may vary from 1000 to 2000 bits in length.

It is likely that a manufacturing system will incorporate several different styles of networking to maximise the transmission characteristics of each network for the particular application. The linear networks resemble data bus lines within the computer itself. The possibility for possible data collision is higher than that for loop networks. The mixing of these different systems is commonly referred to as an 'irregular tree network'.

13.6.4 Network protocols

Logic protocols are used with the physical network topology to prevent data collision on the shared data links. One of the primary factors limiting the efficient development of data networks is the disagreement between transmission standards. Some of the standards in current use are listed in *Table 13.9*.

Table 13.9 Data transmission standards

Standard	*Description*
IEEE-488 (GPIB)	Bus defining physical and data-link control layer; relates to digital communication between computers and various laboratory instrumentation and equipment
RS-232C	EIA standard with several configurations; serial interface between computers and peripheral equipment; uses bit parallel data transmission
RS-366	Standard governing computer-controlled phone dialling
RS-408	EIA standard for numerical control interface
RS-422	EIA standard for balanced voltage digital interface
RS-423	EIA standard for unbalanced voltage digital interface
RS-449	EIA standard for general purpose digital interface; intended improvement to RS-232C
802	Specifies physical and data link layers
802.3	Bus using CSMA/CD method; used in Ethernet applications
802.4	Bus using token passing technique
802.5	Ring (loop) bus using token passing technique
802.6	Bus using MAN access technique

EIA, Electronics Industry Association.

Most of the current network protocol development is intended for local area networks (LAN). A LAN is normally considered to be a network within a single building or close group of buildings. A local area network (LAN) is an intermediate speed data link for restricted distances. Networks with longer data links are classified as wide area networks (WAN). The transmission media required for these long-distance networks are different; however, much of the communication protocol may be the same.

Most current LAN development is based on the International Standards Organization (ISO) Open Systems Interconnection (OSI) model. This model is composed of seven layers which describe the physical and logical nature of the system.

(1) The physical layer describes the physical transmission medium: wires, cables, fibre optics, etc.
(2) The data layer describes the protocol involved with the flow of data in the LAN and ensures correct transmission.
(3) The network layer determines which physical link should be used in a transmission.
(4) The transport layer performs data transfer and error recovery; it ensures reliable transmission.
(5) The session layer provides for access to resources between links.
(6) The presentation layer regulates the appearance and format of the data.
(7) The application layer governs the manner in which programs interact with the network system.

One of the more common hardware connection standards in current use is the RS-232C. This standard was one of the first proposed by the Electronics Industry Association (EIA). It is normally configured as the serial interface between a computer and its peripheral equipment. A typical pin configuration for the RS-232C is shown in *Figure 13.28*. The RS-232C standard has shown several limitations, including distance limitations (normally 50 ft without line loss), transmission speed limitations (20 000 bit/s), configuration inconsistency, limited data bandwidth, cross-talk between data lines, and electromagnetic and radiofrequency (RF) interference.

The EIA RS-449, RS-422A and RS-423A are newer standards intended to alleviate these problems with the RS-232C. These standards specify a 37-pin connector (D-37). There are separate provisions for balanced and unbalanced transmissions.

The IEEE-488 bus is also referred to as the general purpose interface bus (GPIB). This bus has been used extensively in recent years for interfacing to various types of instrumentation.

13.6.5 Typical networks

Typical networks in wide use today are described below.

Arpanet This is a widespread network based throughout the USA. It was originally developed by the US Defense Advanced Research Projects Agency (DARPA). This network consists of over 100 nodes which act as hosts for connection on a regional basis. Each host node is called an 'interface message processor' (IMP) and can serve several connections to the network. Interconnection between IMPs occurs at a variable rate between 50 and 1000 kbit/s. A specialised Network Control Center has been established to monitor failures in communication within this distributed network and can also measure traffic volume throughout the network.

Telenet This is designed similar to Arpanet as a distributed system. There are nine major nodes and hundreds of minor nodes. Many of the local connections use time-sharing operations. Access exists throughout the USA and in many other countries. Telenet is a public, packet switching network.

Prestel This is a network constructed by the British Post Office which uses domestic television sets for local displays and telephone lines for communication outwards to the network. Thus, the input rate to the network (75 bit/s) and the output rate from the network (1000 bit/s) are quite different. Various news agencies and advertisers access the network in order to send information into thousands of homes through a variety of databases.

Euronet This provides information retrieval services to European countries using databases located at several host

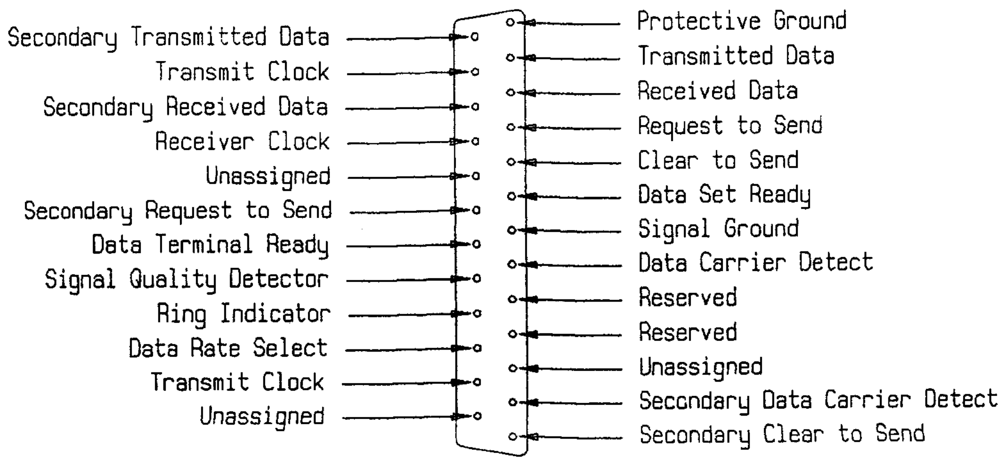


Figure 13.28 EIA RS-232C pin configuration

organisations. This packet-switched network uses 48 kbit/s lines. A French version of this network is named *Transpac.*

Ethernet This is a local area network developed by the Xerox Corporation. It links various nodes within a building or in relatively close proximity via coaxial cables.

In addition, a wide variety of PC based networks have been developed in recent years with the advent of high speed, large storage microcomputers in the 286 and 386 class. Such networks require a large memory unit to serve as a central processor, which is hardwire connected to other units, usually in a star-type network.

13.7 Databases

The term 'database' refers to a collection of stored data which contains some associations or relationships between all or any parts of the data. In general, this collection of data is very large, thus necessitating the use of computer access as opposed to manual data handling. The computer program which controls data input to the database, user access to the database, calculations conducted with data, and output from the database is called a *database management system* (DBMS).

The smallest set of related data which is stored in a database is called a *record.* A record may have very few or many data elements (words, numbers, or lines of information) which are often called *fields*; however, all of these elements or fields should be related in some fashion in order to exist within the same record. An example might be a listing of personnel information for an employee including name, address, job title, salary, social security number, etc. Each item of data occupies a single field within the overall record.

A set of records is called a *file.* In this example, a file would contain all employee personnel records, and might be entitled 'employee personnel data'. The database may contain other files which contain other relevant information regarding the operation of a company. A database may be accessed through a single user environment, such as a PC, or through a multi-user environment such as a network terminal. In the latter case, the database may be shared among several users simultaneously, each performing a different calculation or function with the same set of data.

Databases can be used to generate mailing lists of clients, store personnel records, store and update inventory information, maintain equipment maintenance schedules, handling job cost accounting, and storing look-up tables for various calculations and manufacturing functions. Databases can be operated on a variety of computer types and sizes from mainframes to microcomputers. However, due to the usually large amount of data which is present in a database, it is more appropriate to store the data in a computer with appropriate memory capacity. This would suggest that a minicomputer or mainframe system is most appropriate for database management. However, with proper data communication and/or networking capabilities, it is possible to access a mainframe-based database with a microcomputer or remote terminal.

The database management system is used to provide several key functions. These include

(1) the organisation of data into a variety of requirements as specified by each and every user of the database;
(2) the output of data to a variety of output peripheral devices and in a variety of configurations which can be user specified;
(3) the ability to perform mathematical computations and various sorting functions with all or part of the data; and
(4) the control of database access and sharing among many users.

Because database management systems must provide a variety of features to many users, such software systems are usually quite large and cumbersome. Many such systems require computer programming in special languages that are specific to a given database program. These languages are not usually similar to any standard programming language or operating system. Thus, it may be difficult to learn to develop a database. However, once the database program has been written, it is usually easy to *use* the database, because the user is not aware of the magnitude of the source code or the specific method of database operation. The speed of operation of a database is dependent on the type of database and the number of records which are being accessed.

Database programs may be one of three different categories: (1) hierarchical, (2) network, or (3) relational. A hierarchical database contains a tree-like structure, so that data queries are routed up and down the tree to the final point of selection. This kind of data structuring is very natural and easy to understand. Data queries are accomplished rapidly and the storage of data is managed efficiently in computer memory. However, in order to use a hierarchical database, it is usually necessary to know about the entire tree structure of the data, which may be a burden to the user. The mainframe database manager *Focus* uses a hierarchical database system, as does the micrcomputer version *PC Focus.*

The *network* version of a database management system uses mathematical set theory and is capable of managing very large sets of data. However, this type of database is very difficult to develop, because the relationship between the various data elements is extremely complex. Each set of data has individualised ownership and membership characteristics within the network. This type of database system is difficult to conceive and plan, and is extremely difficult to alter after the system has been developed. However, a network type of database management system is extremely powerful and is well suited for large-scale applications. The *MDBS* series of database programs is the best known of the network database managers.

The *relational* version of database management systems allows concurrent access to multiple data files which are related through common fields of data. Thus, such a system is designed to allow the user to collect and compare data from various records and files. A limitation to the relational database management system is that such comparisons usually require initial data field entry to be consistent among all records and files. The most powerful and well known of the relational database systems is *DBASE*, which has been developed for minicomputers and microcomputers. However, such a system can also be operated in a mainframe computer environment.

The development of database systems for manufacturing and business environments has been widespread. However, such development is an extensive task requiring significant programming experience and proper planning for the implementation and usage of such a system. Often, the selection of a particular database management system will impact the selection of computer hardware and associated interface devices. Thus, the implementation of a database in a manufacturing environment is manpower and hardware intensive, with the development of a database requiring up to 1 year to complete. The actual cost of the database management software is usually the least expensive component of the planning and development process, compared with the programming manpower and hardware purchase components.

13.8 Knowledge-based systems

A knowledge-based system (KBS) is a computer program which stores information relevant to a particular application area and can be used to solve problems in that area in an intelligent manner. Such a system is also called an *expert system* and is sometimes classified under the general heading of *artificial intelligence*. A knowledge-based system contains two major elements: a *knowledge base*, which contains the storehouse of knowledge in a particular area in the form of facts and rules, and an *inference procedure* or *inference engine*, which operates on the knowledge base and performs logical inferences and deductions leading towards a final solution to a posed problem. The use of the term 'expert system' refers to the capacity of a knowledge base to simulate the knowledge of a supposed human expert in the field relevant to the problem at hand. A knowledge-based system with no particular application is called an *expert system shell*.

A knowledge-based system should be oriented to a single application, as opposed to a standard computer language which has more generic functions. A knowledge-based system is 'rule based', meaning that the data are manipulated according to the rules established for a particular application solution (IF...THEN). These rules may be used in an ever-changing order, depending on the information received by the user and interpreted by the knowledge base. This differs from standard computer languages which operate on data in a sequential stepwise manner and is generally 'data driven'. A standard computer program uses easily definable algorithms, while an expert system uses less definable inference techniques.

A knowledge-based system should also contain the following features.

(1) It should be user friendly. The major purpose of a knowledge-based system is to impart expert knowledge towards the solution of a problem which can be used by a non-expert. In fact, its final use is likely intended for a computer novice who may have little understanding of the original design of the software. Thus, the dialogue between the program and the user should allow the user to be directed through the program operation towards a realistic solution of a particular problem. In a manner of speaking, the knowledge base actually gives advice to the user leading towards the solution of a problem.

(2) The method of arriving at a solution in terms of 'how the expert thinks' should be easily interpreted and understood by the user. It is important for the user to believe that the program will arrive at the most logical solution to a problem. The knowledge-based system should be able to describe to the user how the inferences are being made at any point during the program.

(3) The knowledge base should be easily updated, because the knowledge in a particular field will probably advance with time.

Knowledge-based systems usually involve on-line interaction with a user with the aid of a terminal, because the user is often queried continually by the knowledge base in an effort to arrive at the proper deductions. This is similar to a doctor asking a patient questions regarding medical history and symptoms before arriving at a diagnosis.

Specialised computer languages have been created for the development of expert system shells, which can then be used to create specific expert systems or knowledge bases for a particular application area. The most common of these languages are *LISP* and *PROLOG*. These languages involve the use of symbolic processing which uses manipulation of abstract quantities such as words, phrases and terms relevant to the application area in question.[33] Although the terms and expressions in expert system languages are relatively limited and simple, they are quite different from standard computer language expressions and terms. Thus, the development of an expert system usually requires a computer programmer with specific knowledge of expert system development.

In order to implement an expert system, it is necessary to obtain a software system which allows the development of expert system shells, from which particular expert systems can be developed for individual applications. It is important to note that each expert system is specific to a given application area or area of particular expertise. In contrast, the expert system shell is a somewhat generic program that can be used in the creation of many application-specific expert systems.

The use of expert systems usually requires specialised computers in order properly to interpret the rule-based methodology of a knowledge base. Only within the last decade have such hardware and software systems been significantly developed to the point that they have received widespread usage. During this period, expert systems have been used in manufacturing and industrial environments in ever-increasing numbers.[34] Expert systems have been used in the analysis and troubleshooting of electronic circuits, in the selection of appropriate machine parts in an assembly operation, in the analysis of graphic images, and in the planning of facilities design.[34] The costs associated with implementation of an expert system include the likely purchase of a specialised minicomputer or super-microcomputer and associated expert system shell software, and the acquisition of programming expertise that is specific to expert system development.

References

1 RITCHIE, D. M. and THOMPSON, K., The Unix time sharing system, *Commnications of the ACM*, **17**(7), 365–375 (1974)

2 MEALY, G. H., WITT, B. I. and CLARK, W. A., The functional structure of OS/360, *IBM Systems Journal*, **5**(1), 3–51 (1966)

3 NOYES, J., The QWERTY keyboard: a review, *International Journal of Man–Machine Studies*, **18**, 265–281 (1983)

4 SEIBEL, R., Data entry devices and procedures, in H. P.Van Cott and R. G. Kinkade (Eds), *Human Engineering Guide to Equipment Design*, US Government Printing Office, Washington, DC, 311–344 (1972)

5 TAYLOR, R. M. and BERMAN, J. U. F., Ergonomic aspects of aircraft keyboard design: the effects of gloves and sensory feedback on keying performance, *Ergonomics*, **25**, 1109–1123 (1982)

6 NORMAN, D. A. and FISHER, D., Why alphabetic keyboards are not easy to use; keyboard layout doesn't matter, *Human Factors*, **24**, 509–519 (1982)

7 POLARD, D. and COOPER, M. B., The effect of feedback on keying performance, *Applied Ergonomics*, **10**, 194–200 (1979)

8 OHLSON, M., System design considerations for graphics input devices, *Computer*, **11**, 9–18 (1978)

9 RITCHIE, G. J. and TURNER, J. A., Input devices for interactive graphics, *International Journal for Man–Machine Studies*, **7**, 639–660 (1975)

10 SOMERSON, P., The tale of the mouse, *PC Magazine*, **1**(10), 66–71 (1983)

11 WARFIELD, R. W., The new interface technology: an introduction to windows and mice, *Byte*, **8**(12), 218–230 (1983)

12 BURKE, D. and GIBBS, C. N., A comparison of free moving and pressure levers in a positional control system, *Ergonomics*, **8**, 23–29 (1965)

13 ALBERT, A. E., The effect of graphic input devices on performance in a cursor positioning task, *Proceedings of the 26th Annual Meeting of the Human Factors Society*, Seattle, Washington, 54–58 (1982)

14 CARD, S. K., ENGLISH, W. K. and BURR, B. J., Evaluation of mouse, rate controlled isometric joystick, step

keys and text keys for text selection on a CRT, *Ergonomics*, **21**, 601–613 (1978)
15 PFAUTH, M. and PRIEST, J., Person–computer interface using touch screen devices, *Proceedings of the Human Factors Society 25th Annual Meeting*, Santa Monica, California, 500–504 (1981)
16 MIMS, F. M., A few quick pointers, *Computers and Electronics*, 64–117 (May 1984)
17 WHITFIELD, D., BALL, R. G. and BIRD, J. M., Some comparisons of on-display and off-display touch input devices for interaction with computer generated displays, *Ergonomics*, **26**, 1033–1053 (1983)
18 FOLEY, J. D. and WALLACE, V. L., The art of natural graphic man–machine conversation, *Proceedings of IEEE*, **62**, 462–471 (1974)
19 SHEA, C. H. and MILLER, G. E., Eye tracking control for robotic systems, in *Advanced Topics in Manufacturing Technology*, ASME, New York, 39–48 (1987)
20 CALHOUN, G. C., ARBAK, C. L. and BOFF, K. R., Eye controlled switching for crew station design, *Proceedings of the Human Factors Society 28th Annual Meeting*, Santa Monica, California, 258–262 (1984)
21 MILLER, G. E., ETTER, B. D. and SHEA, C., Voice control of manufacturing systems, in *Advanced Topics in Manufacturing Technology*, ASME, New York, 35–38 (1987)
22 DAUTRICH, B. A., RABINER, L. R. and MARTIN, T. B., On the use of filter bank features for isolated word recognition, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1061–1064 (1983)
23 HATON, J. P., Speech recognition in Western Europe, in W. A. Lea (Ed.), *Trends in Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, 512–526 (1979)
24 LEA, W. A., Towards versatile speech communication with computers, *International Journal of Man–Machine Studies*, **2**, 107–155 (1970)
25 DODDINGTON, G. R. and SCHALK, T. B., Speech recognition: turning theory into practice, *IEEE Spectrum*, 26–32 (Sep. 1981)
26 LEA, W. A., Evidence that stressed syllables are the most readily decoded portions of continuous speech, *Journal of the Acoustic Society of America*, **55**, 410A (1973)
27 PICKETT, J. M., Effects of vocal force on the intelligibility of speech sounds, *Journal of the Acoustic Society of America*, **28**, 902–905 (1956)
28 FLANAGAN, J. C., Automatic speech recognition in severe environments, *Final Report of the Committee on Computerized Speech Recognition*, National Research Council, Washington, DC (1984)
29 POLLOCK, I. and PICKETT, J. M., Masking of speech by noise at high sound levels, *Journal of the Acoustic Society of America*, **30**, 127–130 (1958)
30 MARTIN, T. B., Practical applications of voice input to machines, *Proceedings of IEEE*, **64**(4), 487–501 (1976)
31 MARTIN, S. B. and WELCH, J., Practical speech recognizers and some performance evaluation parameters, in W. A. Lea (Ed.), *Trends in Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, 24–38 (1980)
32 WOODARD, J. P. and LEA, W. A., New measures of performance for speech recognition systems, *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, **9**(4), 1–4 (1984)
33 CHADWICK, M. and HANNAH, J., *Expert Systems for Microcomputers*, Tab books, Blue Ridge Summit, PA, 31–44 (1987)
34 KRIZ, J., *Knowledge Based Expert Systems in Industry*, Ellis Horwood/Wiley, Chichester/New York, 11–16 (1987)

Further reading

Computer architecture

GORSLINE, G. W., *Computer Organization Hardware/Software*, Prentice-Hall, Englewood Cliffs, NJ (1980)
HAYES, J. P., *Computer Architecture and Organization*, McGraw-Hill, New York (1978)
LOREN, H., *Introduction to Computer Architecture and Organization*, Wiley, New York (1982)
MYERS, G. J., *Advances in Computer Architecture*, 2nd edn, Wiley, New York (1982)
RALSTON, A. and REILLY, E.D., *Encyclopedia of Computer Science and Engineering*, Van Nostrand Reinhold, New York (1983)
WALKER, R. S., *Understanding Computer Science*, Texas Instruments Corp., Dallas, TX (1984)

Operating systems

PETERSON, J. and SILBERSCHATZ, A., *Operating System Concepts*, Addison Wesley, Amsterdam (1984)
RALSTON, A. and REILLY, E. D. (Eds), *Encyclopedia of Computer Science and Engineering*, Van Nostrand Reinhold, New York (1983)

Computer languages

HELMS, H. L., *Computer Language Reference Guide*, 2nd edn, Sams & Co., Indianapolis (1984)
HOROWITZ, E., *Programming Languages: A Grand Tour*, 3rd edn., Computer Science Press, Rockville, MA (1987)
RALSTON, A. and REILLY, E. D. (Eds), *Encyclopedia of Computer Science and Engineering*, Van Nostrand Reinhold, New York (1983)

Visual displays, keyboards, mice, touch screens, graphics tablets

ANON., 36 High Tech Input devices, *PC Magazine*, **6**(14), 95–202 (1987)
BOFF, K. R. and LINCOLN, J. E. (Eds), *Engineering Data Compendium: Human Performance and Perception*, USAF Armstrong Medical Research Laboratory, Wiley, New York (1988)
RUBINSTEIN, R. and HIRSH, H. M., *The Human Factor: Designing Computer Systems for People*, Digital Press, Burlington, MA (1984)
SALVENDY, G. (Ed.), *Handbook of Human Factors*, Wiley Interscience, New York (1987)
VAN COTT, H. P. and KINDALE, R. G. (Eds), *Human Engineering Guide to Equipment Design*, revised edition, US Government Printing Office, Washington, DC (1963)

Speech recognition systems

ANON., Understanding the Masters Voice, *PC Magazine*, **6**(18), 261–305 (1987)
BRISTOW, G. (Ed.), *Electronic Speech Recognition Techniques, Technology, and Applications*, McGraw-Hill, New York (1986)
LEA, W. A. (Ed.), *Trends in Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ (1980)
LEVINSON, S. E. and LIBERMAIN, M. Y., Speech recognition by computer, *Scientific American*, **4**, 64–76 (1981)

Networks and data communications

DERFLER, F. J., Software connectivity's new frontier, *PC Magazine*, **7**(11), 93–278 (1988)
FRIEND, G. E., FIKE, J. L., BAKER, H. C. and BELLAMY, J. C., *Understanding Data Communications*, Texas Instruments Corp., Dallas, TX (1984)
Local Networks: Distributed Office and Factory Systems, Proceedings of the Localnet 1983 Conference, Online Publications Ltd, Pinner
MAYNE, A. J., *Linked Local Area Networks*, 2nd edn, Wiley, New York (1986)
RORABAUGH, B., *Data Communications and Local Area Networking Handbook*, Tab Books, Blue Ridge Summit, PA (1985)

Databases

ANON., Programmable databases, *PC Magazine*, 7(9), 93–306 (1988)

DATE, C. J., *An Introduction to Database Systems*, Addison Wesley, New York (1976)

DEAKIN, R., *Data Base Primer*, New American Library, New York (1984)

HSIAO, D. K., *Advanced Database Machine Architecture*, Prentice-Hall, Englewood Cliffs, NJ (1983)

RUMBLE, J. R. and HAMPEL, V. E. (Eds), *Database Management in Science and Technology*, North-Holland, Amsterdam (1984)

Expert and knowledge-based systems

BRYANT, N., *Managing Expert Systems*, Wiley, New York (1988)

CHADWICK, M. and HANNAH, J., *Expert Systems for Microcomputers: An Introduction to Artificial Intelligence*, Tab Books, Blue Ridge Summit, PA (1987)

KRIZ, J., *Knowledge-Based Expert Systems in Industry*, Wiley/Horwood, New York/Chichester (1987)

14 Computers in Manufacturing

J Hunt

G M Mair

Contents

14.1 Introduction 14/3

14.2 Computer-integrated manufacturing 14/3
- 14.2.1 The manufacturing organisation and CIM 14/3
- 14.2.2 CIM systems 14/4
- 14.2.3 Selecting a CIM system 14/4

14.3 Manufacturing control systems 14/5
- 14.3.1 Materials requirements planning 14/5
- 14.3.2 Manufacturing resources planning 14/7
- 14.3.3 Capacity requirements planning 14/8
- 14.3.4 Optimised production technology 14/8
- 14.3.5 Shop floor data collection 14/9
- 14.3.6 Computer based simulation techniques 14/11

14.4 Computer-aided design and manufacture 14/12
- 14.4.1 Introduction 14/12
- 14.4.2 CADCAM systems 14/13
- 14.4.3 Mechanical CAD and CAE 14/16
- 14.4.4 Computer-aided manufacture 14/19
- 14.4.5 Electronic CADCAM 14/22
- 14.4.6 Hardware review 14/24

14.5 Numerical engineering and control 14/27
- 14.5.1 Introduction 14/27
- 14.5.2 Numerical control systems 14/28
- 14.5.3 CNC machines—general features 14/30
- 14.5.4 Machining centres 14/34
- 14.5.5 CNC turning centres and lathes 14/39
- 14.5.6 Other CNC machines 14/44
- 14.5.7 Machine controllers 14/46
- 14.5.8 Programming 14/51

14.6 Flexible manufacturing systems 14/58
- 14.6.1 Definition of flexible manufacturing 14/58
- 14.6.2 Advantages and disadvantages of flexible manufacturing 14/60
- 14.6.3 Application areas and development of flexible manufacturing 14/61
- 14.6.4 Elements of flexible manufacturing systems 14/62
- 14.6.5 Work centres 14/62
- 14.6.6 Pallets 14/63
- 14.6.7 Fixtures 14/63
- 14.6.8 Tooling 14/64
- 14.6.9 Transportation methods 14/65
- 14.6.10 Control and planning systems 14/65
- 14.6.11 Simulation of flexible manufacturing systems 14/68
- 14.6.12 Planning a flexible manufacturing system 14/68

14.7 Industrial robotics and automation 14/73
- 14.7.1 Industrial robotics 14/73
- 14.7.2 Industrial vision systems 14/85
- 14.7.3 Automatic identification systems 14/86

14.1 Introduction

Computers are now used in practically all the activities within a manufacturing environment and this chapter covers the major areas of the application of computer systems in the design and production of manufactured goods. The majority of departments in manufacturing companies will have already implemented stand-alone computer systems in support of their own work. Whilst the separate application of computer power has brought about many advantages, it has created 'islands of automation' and has confirmed traditional functional demarcations because the different computer systems and software could not communicate with each other.

The increasing power of computer hardware and software has promoted the development of computer-integrated manufacture (CIM). The aim of CIM is the integration of separate functional systems into a single plant or enterprise-wide system. The greater integration of all manufacturing activities is the key to future competitiveness and even company survival. It is the route that leading companies are already pursuing, and the majority will have to follow to some degree in order to be successful.

In this chapter the term 'manufacturing' is used in relation to the wider scope of the complete manufacturing organisation including the indirect support activities. 'Production' is used in relation to the specific role of the shop-floor departments in making the product.

14.2 Computer-integrated manufacturing

14.2.1 The manufacturing organisation and CIM

'Computer-integrated manufacturing' (CIM) is a general term used to describe the bringing together and integration of the many computer-based activities used in manufacturing. It is not a single technique or application program. The objective of CIM is to control and coordinate the activities of a manufacturing organisation via the use of a single integrated computer system. To do this it must integrate or oversee the many separate computer systems used in manufacturing today.

The separate computer based systems already in major use are materials requirements planning (MRP) and computer-aided design (CAD), engineering (CAE), manufacturing (CAM), test (CAT) and production management (CAPM) applications. Some integration is already occurring as computers and software become more powerful and cheaper. The relationships and development path of separate systems toward full CIM is shown in *Figure 14.1*. Early separate programs are now being developed as modules in integrated suites of manufacturing software. Numerical engineering in the form of CAM is combining with CAD to give CADCAM. Computerised numerical control (CNC) and direct numerical control (DNC) have been combined with planning software to allow the development of the integrated control systems needed to

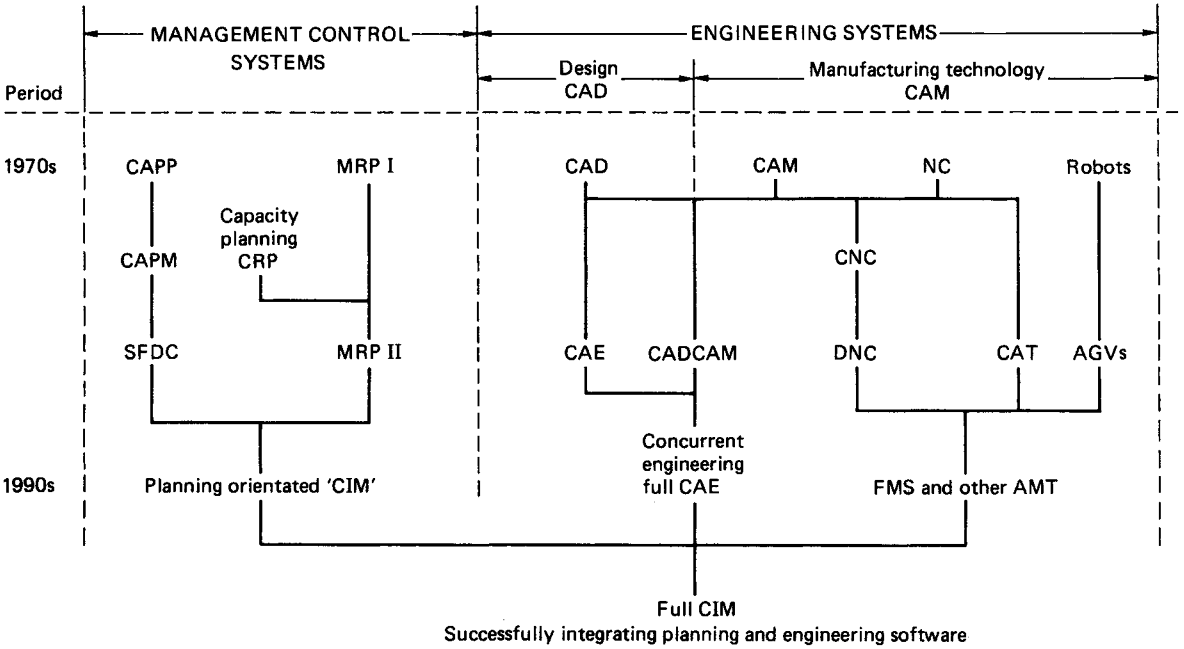


Figure 14.1 Progression of stand-alone computer-aided technologies to full CIM. The figure shows the general application of computer-aided technologies by function and over time. It is a simplified statement and does not show the gradual introduction and continuous development of the technologies over the time period covered.

AGVs	automated guided vehicles
AMT	advanced manufacturing technology
CAD	computer-aided design
CADCAM	computer-aided design and manufacture
CAE	computer-aided engineering
CAM	computer-aided manufacturing
CAPM	computer-aided production management
CAPP	computer-aided process planning
CAT	computer-aided test
CNC	computerised numerical control
CRP	capacity requirements planning
DNC	direct numerical control
FMS	flexible manufacturing systems
MRP I	materials-requirements planning
MRP II	manufacturing-resource planning
NC	numerical control
SFDC	shop floor data collection

run flexible manufacturing installations. MRP, capacity requirements planning (CRP) and other planning programs have been integrated as MRP II.

14.2.2 CIM systems

Computer-aided systems in manufacturing fall into two main groups: manufacturing control systems and engineering systems.

Manufacturing control systems (also known as computer-aided production management (CAPM) or manufacturing information systems (MIS) support the management, planning, scheduling, inventory control and financial functions of a manufacturing unit, i.e.:

(1) manufacturing management controls,
(2) financial controls and budgets,
(3) sales-order control,
(4) production planning and control,
(5) inventory control and purchasing, and
(6) shop-floor management and scheduling.

Engineering systems (computer-aided engineering (CAE) in the widest sense) support the technical functions which process the engineering data relating to the product and its manufacture. Product designs and process control are major features of this software, i.e.:

(1) mechanical product design and engineering analysis (mechanical CAD (MCAD))
(2) electronic product design and engineering analysis (electronic CAD (ECAD)),
(3) manufacturing and process engineering (CAM),
(4) numerical control and technology robotics (CNC/DNC),
(5) quality systems and engineering—automatic testing (CAT), and
(6) software engineering (computer-aided software engineering (CASE)).

Many systems are currently described as being CIM systems, implying wide integration, when in fact they only cover a proportion of the total manufacturing environment—usually within either the management or engineering fields with some limited integration between the two. There are few examples of complete corporate or manufacturing unit-wide CIM systems; this is due to three major reasons.

(1) The necessary powerful and all-embracing integrating software is still being developed.
(2) The cost of such systems is high because the programs are large and complex. In addition, the software ideally needs to be tailored to the particular organisation, products and processes.
(3) The commercial risks involved in changing from a set of separate but functioning computer systems to one central system are very high. Companies are, therefore, cautious of wide-scale CIM implementation on both operational and financial grounds.

There are, however, a number of CIM installations that have proved successful in the integration of the activities within their scope. The subintegration of planning activities on the one hand and engineering activities on the other, has proceeded through the 1980s and produced many successful installations. The key and most difficult integration remaining is that of the management and engineering databases and operational programs.

14.2.3 Selecting a CIM system

A CIM project must be a matter of manufacturing strategy set by senior executives, because the cost and effect on the performance of the company in either success or failure will be significant. Company policies on matters such as attaining world-class manufacturing (WCM) standards, reorganisation into 'flatter' team/product organisation structures, product developments, just-in-time (JIT) working, sourcing policies and future manufacturing locations will have a bearing on the final CIM solution.

The introduction of CIM should be a catalyst for simplifying and restructuring the organisation and operations, as this is likely to bring greater benefits than just the integration, automating or updating of existing practices. When considering the introduction of a CIM system, it must be realised that it will affect every activity and piece of data produced by the operating areas within its scope. This means that every activity and operational requirement should be reviewed.

Check list for a CIM project

(1) Appoint a project team and leader responsible to a senior executive. Set clear objectives for the project scope and team's responsibilities.
(2) Collect a list of desired applications and their advantages from operational managers *and* staff.
(3) Use a series of review teams (having both producers and customers of the data and outputs) to challenge and rank the applications on the basis of:
    (a) data sources and form of input and lead times,
    (b) data destinations and form of output and lead times,
    (c) interrelation of data sets and criticality to each other,
    (d) effect of data loss, corruption or delays,
    (e) security of and access to data, and
    (f) other operational benefits and disadvantages.
(4) Produce an agreed applications list ranked in criticality to the total operation and the costs of success, failure and non-implementation.
(5) Review the labour and skills requirements for both implementation and ongoing support of the new system—operational staff, support staff needed for hardware, software applications and software systems support.
(6) Review additional hardware likely to be needed, and locations.
(7) Review the *practical* lead times for the project implementation and phase out of existing systems—this needs a great deal of care.
(8) Write a system specification based on the above considerations.
(9) Obtain proposals and tenders from a number of system vendors.
(10) Hold a review with the two or three most likely vendors to ensure that they:
    (a) fully understand the specification;
    (b) fully meet the specification for software performance;
    (c) have an adequate specification of the hardware needed;
    (d) have real experience in the specific area of application, i.e. can provide good training for operator, supervisors and support;
    (e) can meet the desired and/or proposed timetable—be particularly concerned with time needed to write/tailor and prove software;

(f) can provide good support during implementation;
(g) can supply sufficient and clear support doumentation;
(h) can provide good support during service at a reasonable cost (N.B. software support is more critical than hardware support).

(11) Visit other vendor installations before selecting the best vendor.
(12) Revise your implementation and running costs on the basis of the vendor reviews in which you will have learnt a great deal of what is really needed.
(13) Only now put forward the expenditure application for final approval.
(14) Agree a contract with non-performance penalties for the vendor in operation and support as well as implementation.
(15) Appoint a joint project team with the vendor and have corresponding members from both sides for your own security and continuity.
(16) Run a pilot scheme to test major assumptions and critical areas.
(17) Hold regular project meetings and reviews for senior management and users.

14.3 Manufacturing control systems

Computer based manufacturing control systems normally include a suite of integrated software modules for the management and planning of production and inventory and the processing of related financial and engineering data. The degree of integration between the modules varies and a prospective purchaser must ensure that the modules offered provide a smooth continuity of operation within the particular manufacturing environment.

The major software applications and modules in use are:

Materials-requirements planning (MRP)
Manufacturing-resource planning (MRP II)
Capacity requirements planning (CRP)
Optimum production technology (OPT)
Shop floor data collection (SFDC)
Simulation programs

14.3.1 Materials requirements planning

14.3.1.1 Background

MRP has been the major module used in computer based manufacturing control over the last 20 years. The data sets needed to control the material flows of a manufacturing unit are fundamental to the operation of the unit and the data are produced or used by all the core departments in the unit. Hence a MRP system will occupy a pivotal position in the control and planning of manufacturing. This role is illustrated in *Figure 14.2*.

Even a small company or manufacturing plant may have over 5000 part numbers to handle, and large units may have to deal with 50 000 part numbers. This represents a large, constant workload in deciding, purchasing and controlling the flow of production materials in a factory. A large amount of data must be processed for each part number on a regular basis, creating a significant data processing workload. This task, which is common to every manufacturing unit, was an obvious early and commercial candidate for automation on a computer.

The programming techniques developed to handle this task are called materials requirements planning (MRP). This is now sometimes called MRP I to distinguish it from a later related, but broader, development called 'manufacturing resource planning' (MRP II) (see Section 14.3.2).

Because of the large amount of data processing and memory needed for MRP, most early systems were installed on the company or factory mainframe or mini computer which had already been installed for financial and sales work. With the dramatic increase in the processing power of supermicros (the 80386 and 80486 machines) the reduction in memory costs and the introduction of cheap local area networks (LAN), it is now possible to run all but the largest MRP programs on an appropriate desk-top hardware system.

14.3.1.2 System operation

Types of data needed Two types of data are needed: static and dynamic data.

Static data are data that are standard for some time. Such data comprise:

(1) Details of the product structure, parts lists and material specifications, as produced by the design office. These are entered into the MRP database either manually or, increasingly, automatically from CAD.
(2) Order rules and production and purchasing lead times for each part number.
(3) Shop-floor department numbers and material routings.
(4) Factory standard costs and times; purchase costs.
(5) Stock location numbers in stores (if an appropriate location system is used).

Dynamic data are data that change with sales demand and production ouput.

(1) A master production schedule (MPS) is entered from production control. This specifies the gross output required and is derived from a sales forecast that normally looks at least 3 months forward and often up to 1 year. In some cases the sales forecast may be taken directly.
(2) Stock levels—opening, closing, issued, committed, free, safety, production, spares, repairs.
(3) Completions—MRP will be sent reports on stores issues against requisitions for line stock or JIT, production completions, stores receipts, purchases receipts, work in progress (WIP) and scrap.

Calculations A new MPS gross requirement is entered periodically—usually weekly, bi-weekly or monthly, depending on the nature of the business. A MRP run is a large and lengthy computer operation and produces a large number of printed reports. Thus the frequency of the run must be the optimum for the rate of sales-demand changes, the response flexibility of production, the capacity of people and the system to handle large amounts of changing detail and the costs of doing so. The system calculates the nett requirements by taking account of total stock, stock already committed to orders, safety stock levels, scrap allowances and the resultant free stock quantities. Production and material purchase lead times are applied to arrive at completion due-dates and order-release dates.

Outputs A materials-requirements report is issued, normally to inventory control, purchasing and production control departments. Purchasing recommendation reports are issued to the purchasing department who will review the situation and initiate the release of purchase orders—either producing them manually or sanctioning the release from a computer based purchasing module linked to MRP.

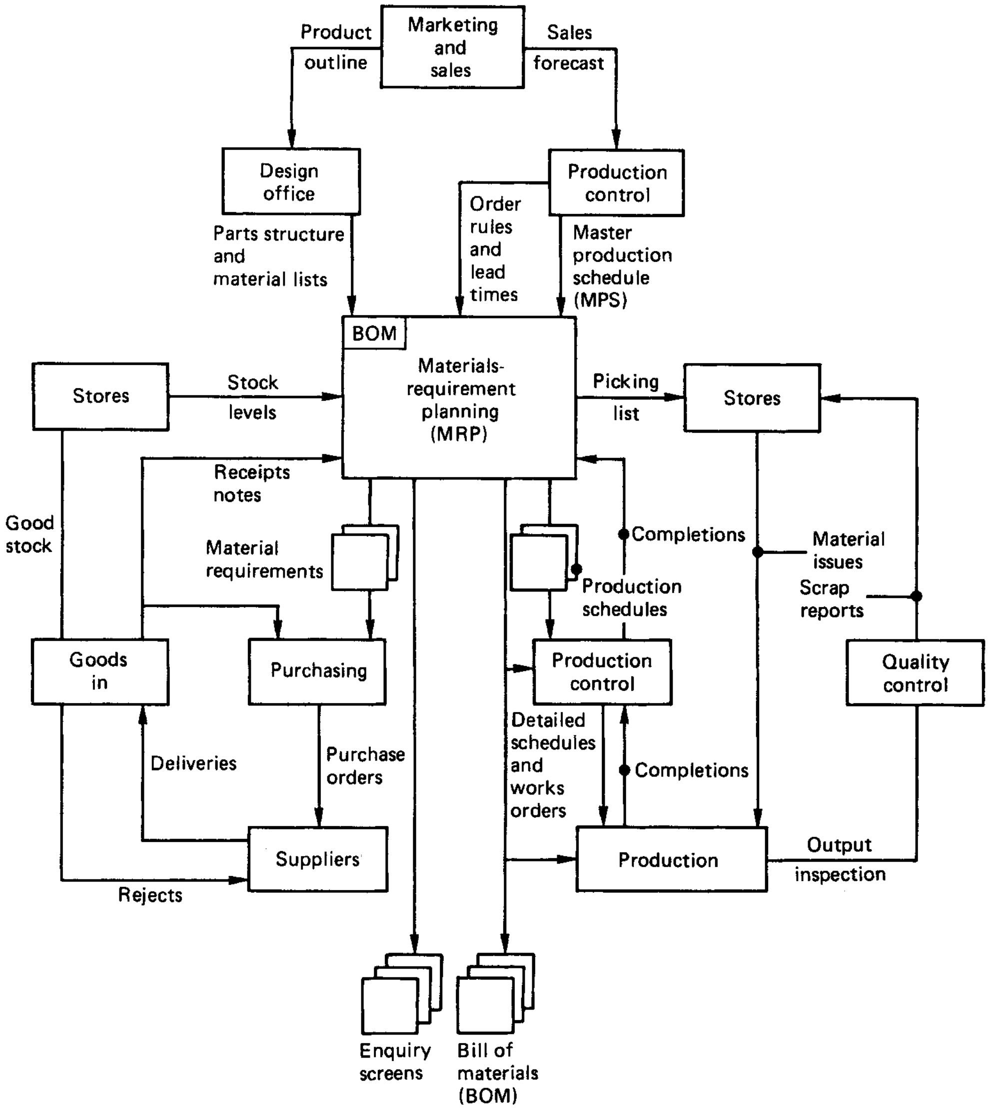


Figure 14.2 The pivotal role of MRP in the control and planning of manufacturing

Stock picking lists will be issued to the various stores. The actual quantity picked and shortfalls will be fed back.

A bill of materials (BOM) for each product and subassembly is produced. This reflects the order and method of production, which is usually different in structure from the parts list produced by the design office. This difference should decrease as the drive to closer integration and design for manufacture techniques takes effect.

Stock status, inventory value and many other reports are normally produced.

Data screens There is a constant demand for instant access to data (particularly stock statuses and inventory costs) from a number of departments, i.e. inventory control, purchasing, production control, production, manufacturing engineering, design and finance. Consequently, it is necessary to have data available at terminal monitors in these departments. The data will be set out on a number of 'screens' that detail specific areas of data. The screens are related in a formal structure and are usually accessed via a menu and 'next screen' facility. A typical stock status screen is shown in *Figure 14.3*.

Personal access to read most data is normally unrestricted or moderately restricted by password because a large number of people need access. Conversely, the ability to modify data will be strictly controlled and only possible at certain terminals—normally within the 'owning' departments of inventory control and/or production control. Terminals will be connected to the computer holding the processing and central memory by a network. Most read-only outlets will be 'dumb terminals' at least insofar as the MRP application is concerned.

Part No.	abc12345			Date	11/05/92
Revision	3			Time	15:03:00
Effective	10/4/92				

Used In	xyz5677	xyz5678	xyz5679
Qty per part	4	4	2

			p
Material Part No.	AL6789	Labour cost	24.123
Revision	2	Material	10.123
Effective	20/10/89	Overhead	5.123
Annual demand	8000	Standard cost	39.369

Stock status

Total		2000	
Committed – Production	1000		
– Spares	200		
– Repairs	100		
Total		1300	
Free		700	
On order		1000	due 12/6/92
Covered		1700	

Demand

Period—months	5	6	7	8	9	10
Qty	500	600	1000	900	800	800
Covered	500	600	600			
To cover			400	900	800	800

Issues

	Qty	Date	Order No.
This period 1	100	10/4/92	456789
2	200	20/4/92	456789
3			
4			
Last period	600		
Year to date	2400		

Figure 14.3 Typical stock status screen

14.3.2 Manufacturing resources planning (MRP II)

Materials requirements planning (MRP) systems have a major limitation in that, having determined the demand in terms of part numbers, volumes and material required, they do not address the next major issue, i.e. the capacity of the factory to meet the demand. If there is a significant mismatch between the two factors, then MRP will produce an unobtainable target, which will result in wasted time and overstocking as the factory capacity fails to convert materials and parts into finished goods.

MRP II addresses the subject of capacity and financial planning and integrates the whole manufacturing control and planning environment from sales forecast to shop-floor control. In this sense it can be regarded as a major component of a CIM environment. The major elements of a MRP II system are shown in *Figure 14.4* and a hardware environment is shown in *Figure 14.5*. It can be seen that the MRP holds a central position, but the computer-based controls and planning are extended to cover:

(1) sales-order management,
(2) purchase-order management,
(3) works-order management,
(4) capacity requirements planning,
(5) shop-floor control, and
(6) cost accounting.

These are present in the form of software modules that are integrated with the necessary links to each other. The system has the capability of accepting a sales or business plan and developing a master production schedule from it. This would then drive a rough cut capacity plan to see if capacity matched demand. If the capacity is regarded as finite and is not sufficient, the original sales/business plan can be modified. Once a master production schedule has been accepted, the MRP and production control modules can proceed.

Surveys have shown that MRP II has met with some initial problems in a number of installations. There are a number of reasons for this, but one of the most common is the difficulties met in organisations having to absorb such a wide ranging

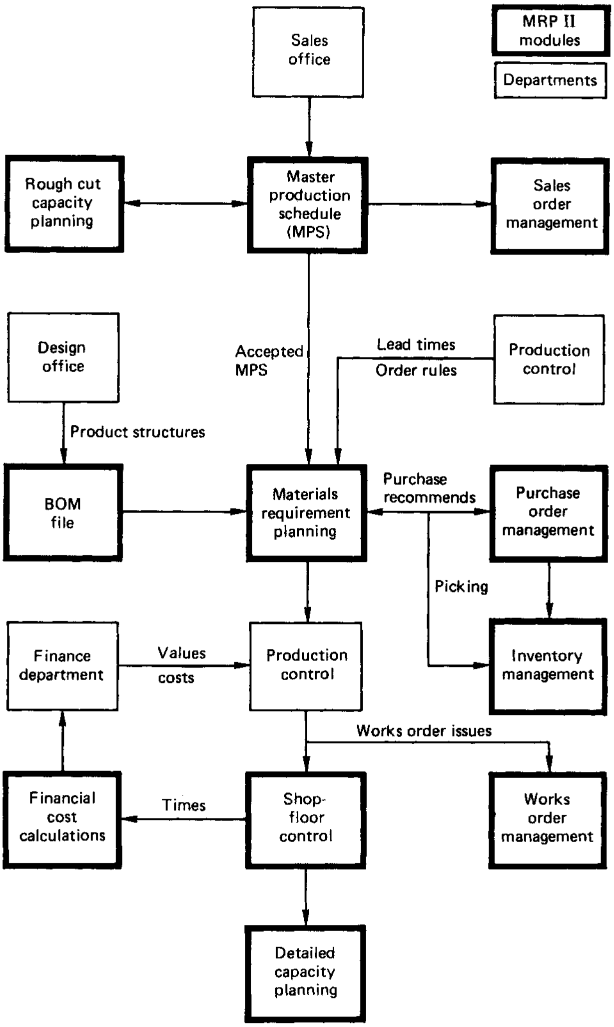


Figure 14.4 A MRP II structure. BOM, bill of materials

system and the many changes it enforces. Secondly, there has been (as with so many major projects) an underestimation of the time needed for successful implementation. Thirdly a reluctance by some customers to spend sufficient staff time on training in the new systems has decreased their success rate.

A major disadvantage with MRP II systems is that they are expensive because of their size and wide scope. The structure of these systems and their cost tends to retain centralisation of activities and decision-making in the traditional service departments. This runs contrary to current manufacturing philosophy which is to product-oriented working groups (i.e. group technology for a complete product and its support services) and 'flatter' structures. In this environment the decision-making and support services are being devolved to the semiautonomous product groups. Also the use of JIT is removing the need for in-depth capacity planning on a large-batch basis.

14.3.3 Capacity requirements planning

Some planning systems, including MRP, assume that infinite capacity is available. This is clearly not so, and at some point capacity requirements must be taken into account. This can be done by using capacity requirements planning (CRP) which is a module within MRP II and can also be used in conjunction with MRP.

CRP modules need data on the process routings, sequence, set-up times and process times in order to calculate the capacity situation. Advanced programs provide simulation and a 'what if' capability for capacity planning.

14.3.4 Optimised production technology

Optimised production technology (OPT), is both a measurement philosophy for managing a manufacturing unit and a proprietary scheduling technique.

14.3.4.1 Philosphy

The 'philosophy' is a method of defining the goals and measuring the performance of manufacturing units. It uses three parameters which, it is claimed, are more easily understood than traditional financial measurements. The OPT measurements are said to be more readily identifiable with the global aims of a manufacturing unit.

Traditional measure	*OPT measure*
Net profit	Throughput
Retain on investment (ROI)	Inventory
Cash flow	Operating expense

Throughput is the rate at which the manufacturing unit generates money from actual sales—it does not value production output (which would not highlight finished stock still held as inventory).

Inventory is a measure of raw material value only. Added value of manufacturing appears as operating expenses. Finished production unsold is also included.

Operating expense is the money spent to convert inventory (raw material) into throughput (sales revenue).

The goal of OPT is to increase throughput (sales) whilst decreasing inventory and operating expense. Supporters of the method believe that the measures give a clearer understanding of the overall goals of a manufacturing plant through the visibility of the measurement of conversion of raw materials to sales at minimum cost—something that can be lost in the use of traditional accounting methods. Ongoing improvements can thus be promoted and implemented more clearly when their performance is measured in this way.

OPT involves building a mathematical model of the factory from which simulation and scheduling can be created.

14.3.4.2 OPT scheduling

The OPT scheduling method is an alternative to the demand and due-date backward scheduling of CRP and MRP II. There are a number of rules to be followed, some of the most fundamental being as follows.

(1) Production flow not production capacity should be balanced.
(2) Capacities should be identified as 'constraining' or 'non-constraining' because the constrained resource will directly affect throughput (sales) performance. Con-

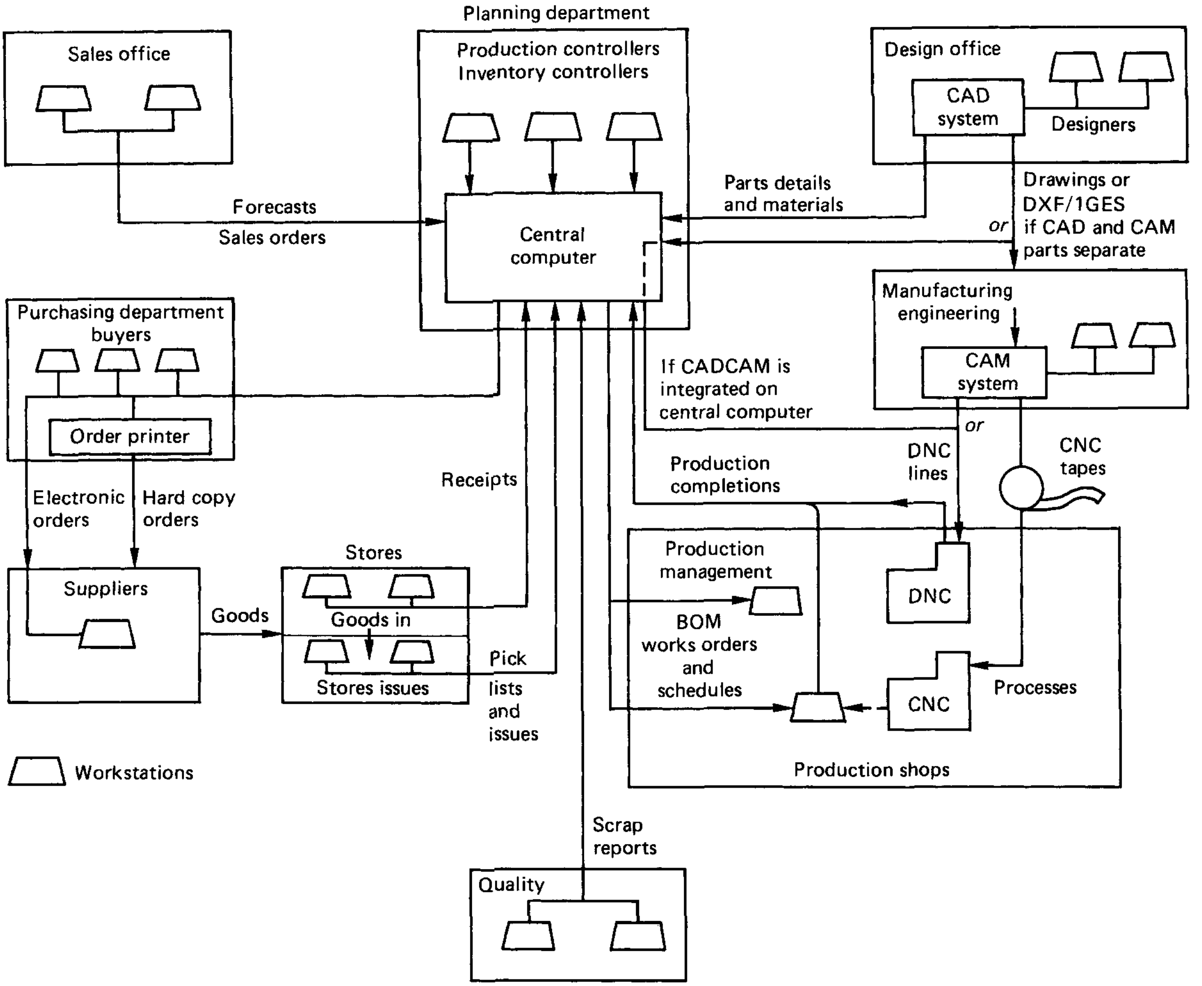


Figure 14.5 Hardware relationships in a MRP II system (financial lines excluded). BOM, bill of materials; CAD computer-aided design; CAM, computer-aided manufacturing; CNC, computerised numerical control; DNC, direct numerical control

strained resources are forward scheduled to their finite capacity, whilst non-constrained ones are backwards scheduled to infinite capacity. Any buffers required are expressed in terms of time, not inventory.

(3) There are two types of batch quantity—transfer batch sizes and process batch sizes. By balancing flow through balanced transfer batch sizes, lead times can be reduced. This is shown in *Figure 14.6*.

Because OPT balances flow it tends to work on smaller batch sizes. This can bring large reductions in inventory and lead times but, like JIT, it requires a new attitude and commitment to the philosophy from staff—in the face of the traditional seemingly more efficient production obtainable from large batch sizes. OPT can be used in conjunction with MRP II and JIT.

The functions of the OPT system are shown in *Figure 14.7*.

14.3.5 Shop floor data collection

Shop floor data collection (SFDC) provides an efficient method of collecting data from a number of locations around a factory and feeding them directly to a central computer where they are processed. The types of information collected include:

(1) production batch completions—codes and quantities;
(2) job time booking—process, set up, waiting, breakdown;
(3) staff attendance and time booking;
(4) quality reports—SPC;
(5) work in progress and stock recording;
(6) tooling and engineering records;
(7) machine output (direct count from machine); and
(8) machine time logs.

Programs normally allow the user to design report and/or screen layouts.

Data are collected by using:

(1) real-time direct machine-to-computer communication;
(2) data-input terminals;
(3) a combination of (1) and (2);
(4) the input devices are connected to the central computer by a local area network (LAN).

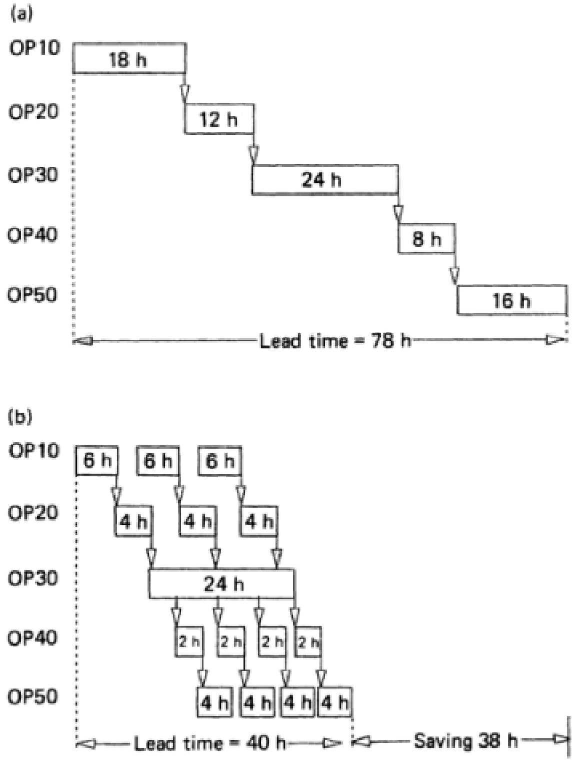


Figure 14.6 Predictions of lead time from OPT. (a) Transfer batch equal to process batch. (b) Transfer batch not equal to process batch

14.3.5.1 Input devices

Shop-floor terminals These are specially designed units which are simplified and ruggedised for shop-floor use. They have a small keyboard and display which allow staff to key in data. Some units have a card-reading facility which sends identification cards that are similar to credit cards.

Personal computers Personal computers (PCs) are used for input and also as monitoring devices to give the shop floor and supervision a view and feedback of the data being transmitted. In many cases the 'PC' is a remote terminal to the central computer. It will have a monitor screen and keyboard, but no processing or memory capabilities.

Bar-code readers These read bar codes put on control documentation and products. Their major advantage is the elimination of the keying errors and speed.

Machine data units These are fitted to machines to automatically capture data on output and other features of the machine utilisation.

14.3.5.2 Central computer system

In some applications, data are fed to the main planning computer (a mainframe, mini or supermicro). In other more independent data collection systems, a PC is used as a file server or the actual computer. A 286 processor and 40 Mbyte hard disk are usually sufficient for systems having less than 100 input points.

Data are transmitted from the input points to the computer via a local area network (LAN).

A SFDC system is shown in *Figure 14.8*.

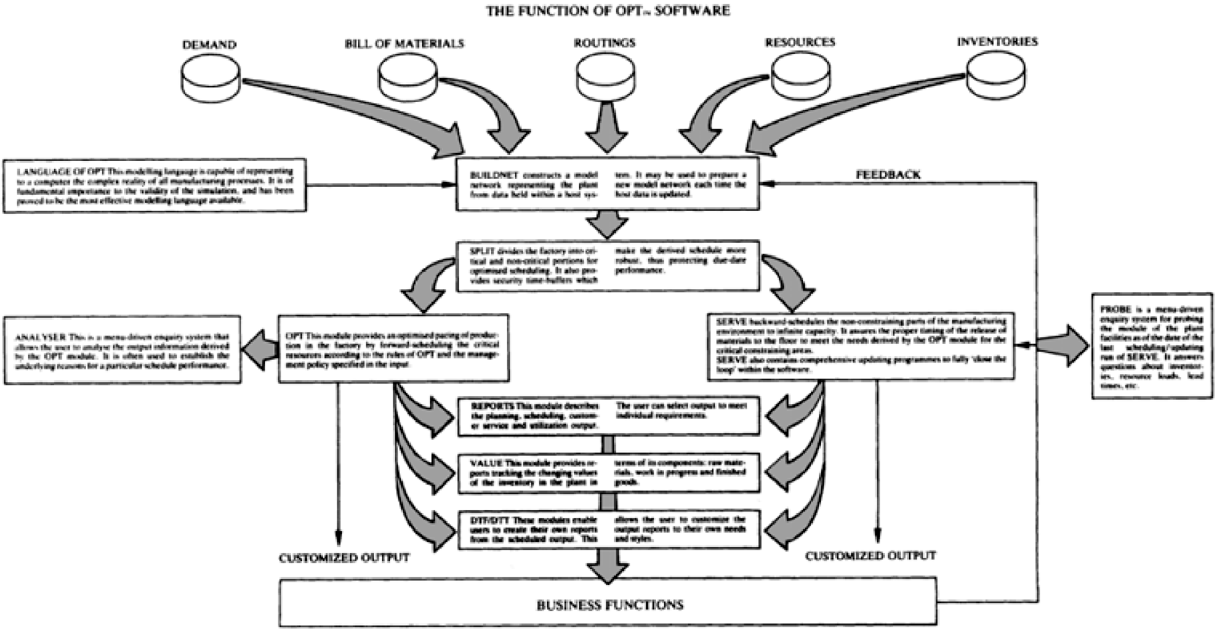


Figure 14.7 The function of OPT software. (Courtesy of Timegate Manufacturing Solutions)

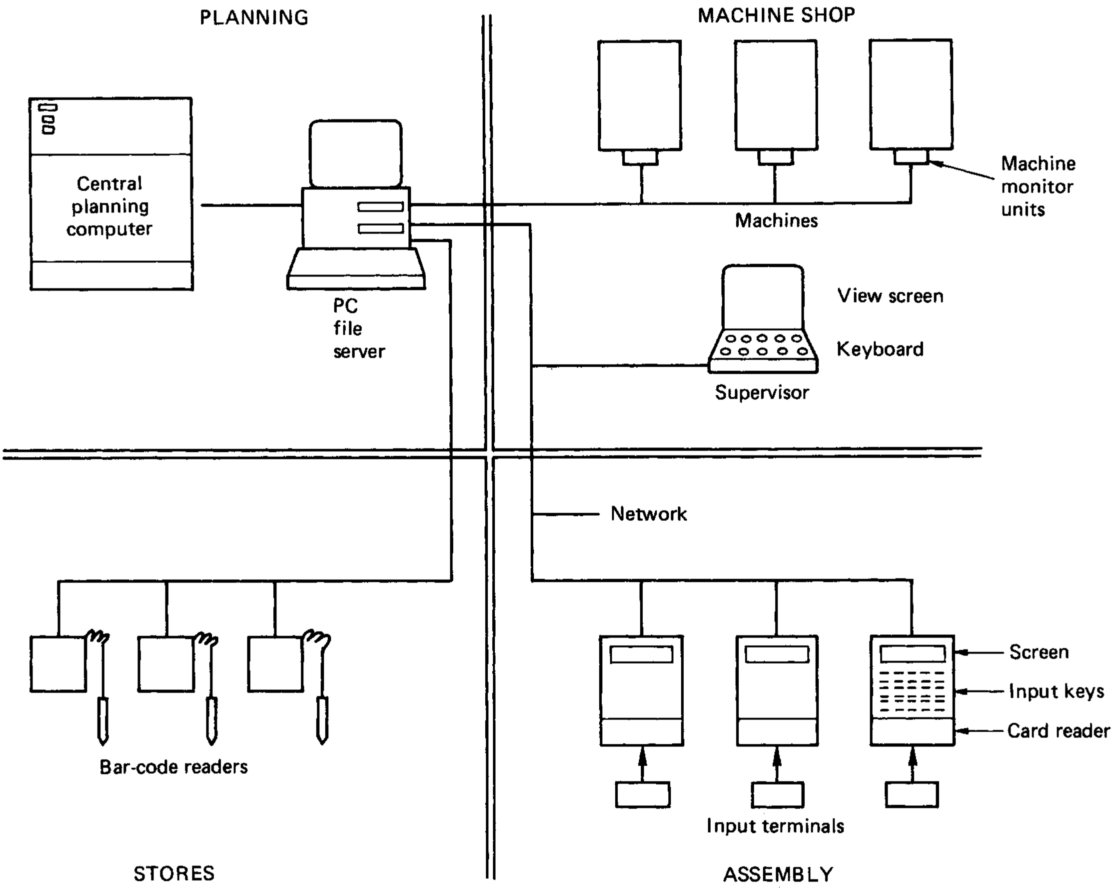


Figure 14.8 Elements of a SFDC system

14.3.6 Computer based simulation techniques

14.3.6.1 Introduction

Manufacturing systems are complex and dynamic. They combine a large number of activities, facilities and objectives in order to achieve the design and making of a product. Their complexity is compounded by the fact that the systems are dynamic—conditions within them are constantly changing as orders are received, plans initiated, materials flow and tasks completed.

The use of computer-based simulation is growing because it is a tool that allows detailed forecasting which was previously impossible to carry out by manual methods, due to the amount of data to be processed in even a simple system. The ability of a computer to hold a model of a system as a series of mathematical algorithms and its ability to store and process vast amounts of detailed data, now makes simulation of systems a practical possibility.

Simulation is normally carried out by a module within an application program. A fundamental requirement of simulation is the construction of a mathematical model of the system under consideration. The mathematical model can then compute likely outcomes from the input conditions specified by the user. Areas in which simulation is used include:

(1) production control and management,
(2) MRP II,
(3) OTP scheduling,
(4) factory planning and layout,
(5) flexible manufacturing systems (FMS) design,
(6) computer-aided engineering (CAE),
(7) computer-aided design (CAD), and
(8) computer-aided manufacture (CAM).

The output of simulation can be as text, but is increasingly graphical for ease of understanding. *Figure 14.9* shows a graphic simulation from a factory planning program.

14.3.6.2 Major elements of simulation

'Simulation' is defined in the *Oxford English Dictionary* as: 'the technique of imitating the behaviour of some situation or system (economic, military, mechanical, etc.) by means of an analogous situation, model or apparatus, either to gain information more conveniently or to train personnel'.

As far as FMS design is concerned this may be re-expressed as: 'a technique that assists design by testing a model of the proposed system with the range of likely operational conditions'.

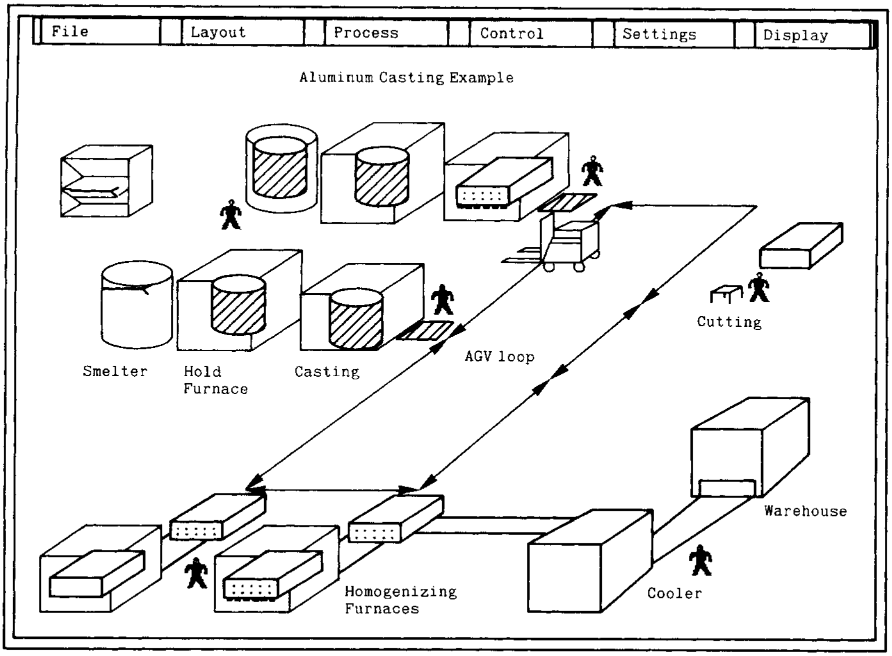


Figure 14.9 Factory planning simulation graphic of a factory. (Courtesy of CACI Products)

A 'model' is defined as: 'a simplified or idealised description of a system, situation or process, often in mathematical terms, devised to facilitate calculations and predictions'. Mathematical models will demonstrate the effect of the operational inputs, outputs and loads on the system, and this is the simulation technique that is discussed here.

Types of simulation There are three basic types of simulation, all of which use the technique of recording changes in the state of a model at points in time throughout a defined period. The method of establishing and using the points in time is one of the major differences between the three types. The three types are described below.

(1) *Discrete*—assumes that changes only occur at distinct points in time and that they occur instantaneously. No changes take place between points. Manufacturing simulation software packages are based on this method.
(2) *Continuous*—assumes that changes are continuous throughout the period. The simulation of continuous material flows, levels of materials and the application of mechanical forces are typical applications.
(3) *Combined*—includes both discrete and continuous simulations. This is the more common situation in real-life systems where many activities of either type impinge upon each other.

14.4 Computer-aided design and manufacture

14.4.1 Introduction

The development of CAD, CAM and CADCAM has made a great impact on design and manufacturing and now offers a wide choice of software applications and systems hardware. A number of terms are used to describe the fields of application and are generally taken to mean the following.

CAE (computer-aided engineering)—(1) in the widest sense the term covers all engineering activities that are run on computers, e.g. CAD, CAM, CAT and CASE. (2) In the narrower sense, it encompasses design and analytical activities such as finite element analysis (FEA), and dynamic simulation programs, but does not include computer-aided design or draughting.

CAD (computer-aided design)—this term is commonly used to describe most computer-aided design and draughting facilities. Two-dimensional CAD is predominantly a draughting medium with some limited design capabilities. Three-dimensional CAD provides enhanced draughting plus geometric modelling.

CADD (computer-aided design and draughting)—this is a fuller description of the normal design and draughting facilities, but as there is now a general trend to use 'CAD' when meaning 'CADD', the latter term is becoming redundant.

CAM (computer-aided manufacture)—(1) In the widest sense this encompasses all modern manufacturing technologies that use computers in a central role. Thus all CNC and DNC machines, robots and flexible manufacturing systems are elements of CAM. (2) In the narrower sense, CAM refers to the computer-aided part programming systems for CNC/DNC machines. An in-built or linked CAD facility is used to create the part geometry prior to programming and can also be used to design tooling and fixtures. Modern CAM software also has production planning and scheduling capabilities to assist in the control of production on the programmed machines.

CADCAM (computer-aided design and computer-aided manufacture)—these are systems in which the traditional CAD and

CAM elements are integrated. The CAD and CAM programs share a common database, which allows the part geometry needed for CAM can be taken directly from the drawing files. Parts can be machined under the direct control of the software (via DNC) once the design is established.

CAE has been widely applied to aeronautical, civil, chemical, electrical and electronic manufacturing and mechanical engineering. The applications that the manufacturing engineer is most likley to encounter are CAD, CAM and CADCAM systems used in the design and manufacture of mechanical and electrical/electronic products.

Figure 14.10 shows the relationships between the modules commonly used within CAE, CAD and CAM.

14.4.2 CAD/CAM systems

14.4.2.1 System configurations

Centralised systems (Figure 14.11(a)) These employ one central computer (a 'host computer') which carries out all the processing, and controls all the terminals and other hardware elements of the system. This has the advantage of there only being the need to maintain one computer but, if the computer also serves other functions such as finance and production, there are likely to be operational delays because CADCAM is a heavy user of processing power and memory. The ensuing slower response time is likely to be unacceptable for the interactive nature of CADCAM work. A host computer dedicated to CADCAM is clearly a better operational proposition than is a share facility, always assuming that it is justified economically.

Multiple-host systems (Figure 14.11(b)) A number of host computers are used to service groups of terminals with most of the peripherals connected to one host. This gives a faster response because the processing capacity is increased and more localised. This provides additional security over having a number of host computers. The hosts are usually 32-bit minicomputers the individual cost of which will be lower than the required single mainframe host, but collectively the cost may be about the same. Multiple host systems may be the best solution to a functional or geographical split in the total CADCAM environment.

Distributed systems (Figure 14.11(c)) In this case the processing and random access memory (RAM) and read-only memory (ROM) capacities are distributed to the terminal

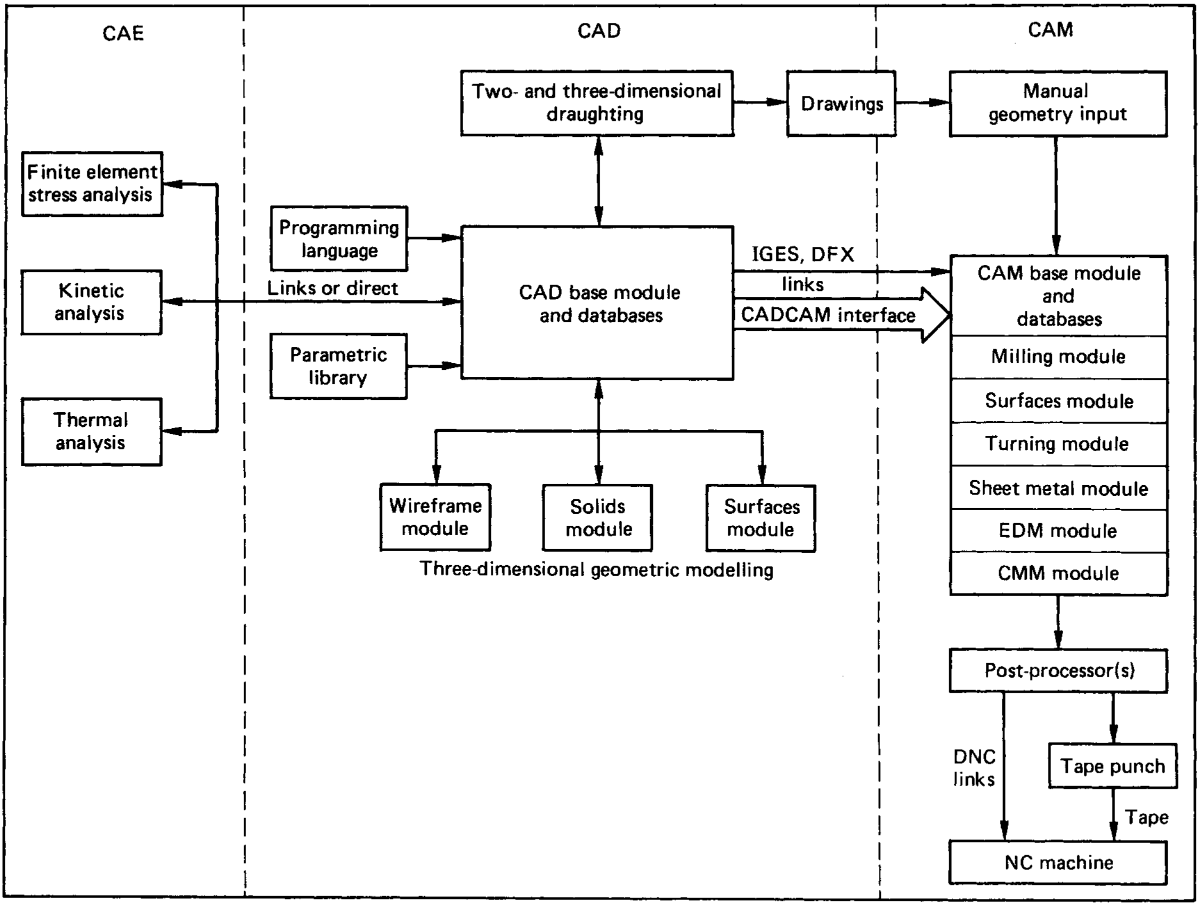


Figure 14.10 Relationships between mechanical CAE, CAD and CAM. EDM, electrical discharge machining; CMM, coordinate measuring machines

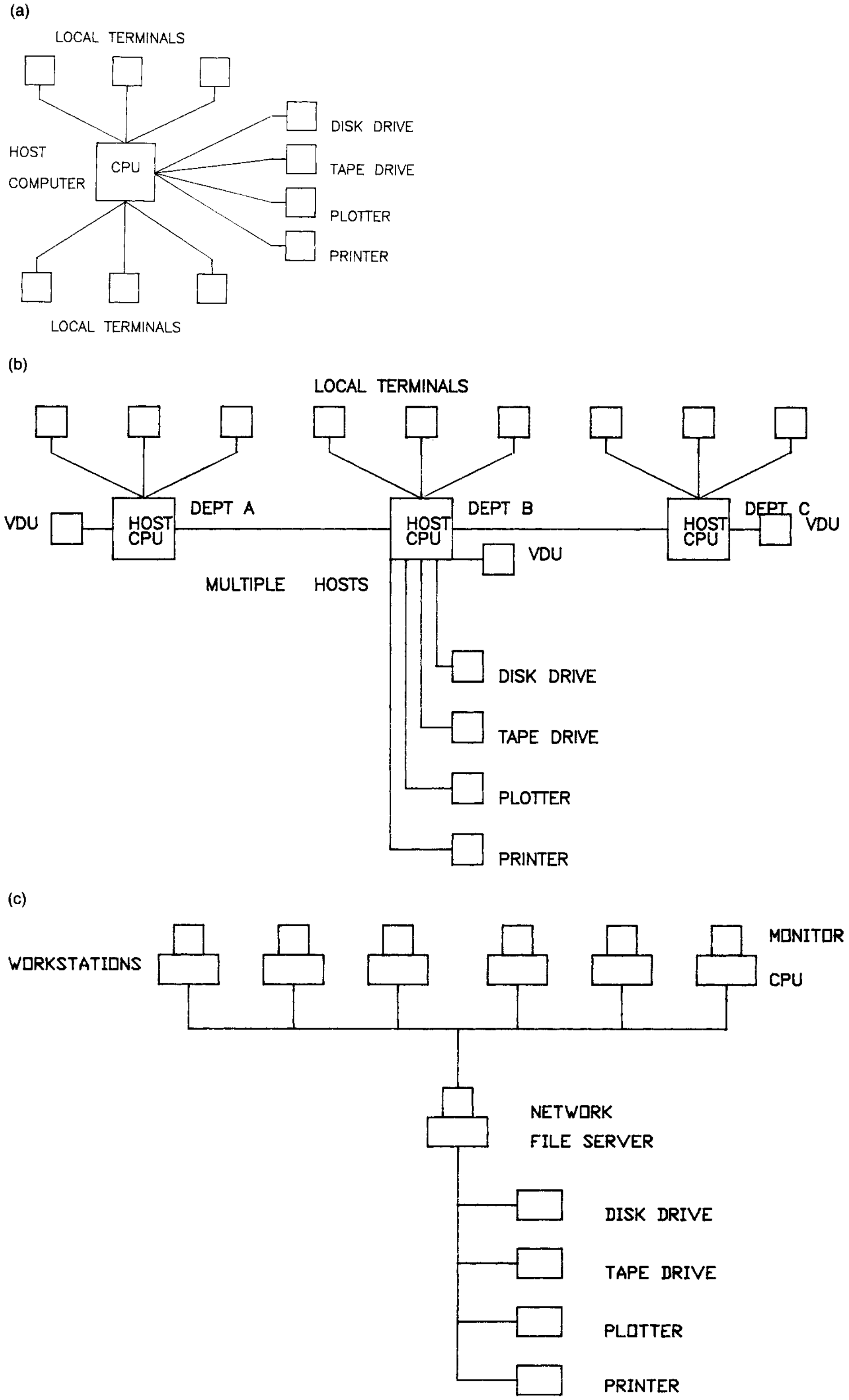


Figure 14.11 System configurations. (a) Centralised system. (b) Multiple-host system. (c) Distributed system via a local area network. CPU, central processing unit

positions by replacing the 'dumb' terminals with intelligent workstations that are connected by a local area network (LAN). In order to control the communication of the workstations with the shared peripherals, a computer known as a 'fileserver' is also needed. This carries out the communications and network control functions of central hosts, but leaves the applications processing at the individual workstations. Response times are kept to a minimum because each workstation has its own processor.

Stand-alone (single-workstation) systems This is the basic configuration in which a single computer is connected to its own plotter, printer and other peripherals. These may be mainframe or mini computers but a very large proportion are PCs. Stand-alone systems are widely used in industry and professional engineering because they have been bought to satisfy an immediate demand of individuals or small departments. However, with the general increase in the number of workstations being installed there is a growing trend to connect these into a local area network (LAN), thereby creating a distributed system for sharing plotters, printers, discs and other peripherals.

14.4.2.2 System hardware

A wide range of hardware has been developed in order to replicate the resources and capabilities of the traditional design, draughting and manufacturing engineering functions. The combinations of hardware used will depend on the system configuration. A typical CADCAM set-up is shown in *Figure 14.12* and the individual types of hardware are discussed in more detail in Section 14.4.6.

Host computers These are central computers that provide the processing, memory and system communications and controls with the terminals acting only as input/output devices. All printers, plotters, tape punches and other peripherals are controlled by the host.

Terminals These provide the point of interface between the engineer and a host computer. Usually comprising a monitor, keyboard and other input devices, they have no or limited processing and memory capacity.

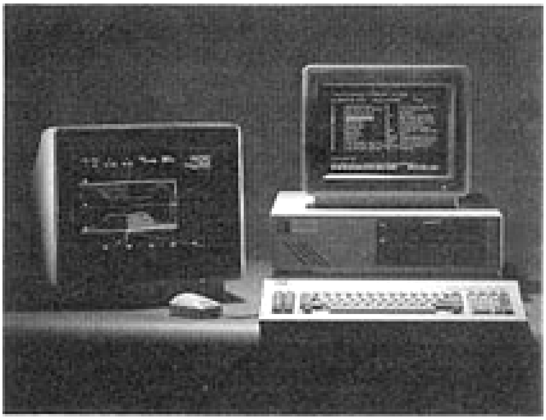


Figure 14.12 Typical PC-based CADCAM system with two screens, a keyboard and a mouse. (Courtesy of Pathtrace Engineering Systems)

Workstations These are 'terminals' that are computers with processing and memory capabilities. The engineer or designer works at a graphics monitor using a variety of input methods such as keyboard, digitising tablet, mouse, tracker ball or light pen. The workstation may be a purpose-designed unit or a PC with sufficient processing power and memory. A workstation may have one monitor, but more usually has two—a full-colour high-resolution screen for graphics work and a second (often mono) screen for the display of commands, data listings and system management activities. Some systems provide two high-resolution graphics screens to allow parallel working in two levels or views.

Fileserver This is a computer in a networked (distributed) system that is used to supervise the communications between workstations and the shared peripherals. On some systems it will also hold the file mass storage capacity.

Mass storage CAD and CAM systems need a large amount of memory for long-term storage of designs and data. This usually takes the form of hard-disk drives on most machines. Tape storage was the first medium to be used (on mainframes) but is being superseded by disks which have much shorter access times and, being sealed units, are less susceptible to damage.

Plotters These produce drawings automatically by a variety of technologies. They are usually shared by workstations via a network. Standard sizes range from A4 to A0. Larger plotters are available for special applications.

Printers These produce hard copy of programs and data normally on A4 or A3 size paper.

Tape streamer This is used to back up all files onto a tape at regular intervals (every few hours, or daily), in order to avoid the complete loss of files or the work of the current period due to an accident of some sort. Tape is acceptable in this case because real-time fast access is not a requirement.

Tape punch This is used for the production of paper tapes for numerical control machines.

14.4.2.3 Software

In common with normal computing practice, three types of software are used:

(1) operating systems software which controls the functions of the hardware;
(2) communications software which provides the communication between the hardware; and
(3) applications software which carries out the specific task required by the user.

Operating systems software Some of the most widely used operating systems are:

(1) *VMS*—the DEC (Digital Equipment Corporation) operating system for the VAX model of minicomputers which are a major hardware platform for CADCAM systems.
(2) *UNIX*—a hardware-independent, multitasking system developed by Bell Laboratories, which has been applied to many platforms. UNIX has many variants such as XENIX, ULTRIX, HP-UX and Sun OS which are supplied by other vendors.
(3) *DOS*—the 'disk operating system' developed by Microsoft for stand-alone PCs. It has the disadvantage of being

a single-task system which is unsuitable for multiuser or multitasking operatings.

The operating system that seems destined to have become the most universally used is UNIX and its variants, because of their machine independence, multitasking and multiuser applications plus the already established wide usage which makes it a *de facto* 'industry standard'. All the major CAD applications packages have been written to work on UNIX machines. With the general move towards more 'open' (e.g. hardware-independent) systems, modern PCs are being supplied with both DOS and UNIX type operating systems.

Communication software (protocols) With the increasing emphasis on CIM and open systems, it is important that different computer systems can communicate, i.e. are capable of securely transferring data and files between one another. In order to do this they must have a common software protocol to carry out the transfer. Communication protocols in wide use include DXF (Data Exchange File) from AutoCAD, and IGES (Initial Graphics Exchange Specification). Most system suppliers provide these links plus other protocols for communication with various widely used systems.

Languages used The majority of early CAD software was written in FORTRAN because this was the primary language used in engineering and scientific fields. Many modern packages are now being written in C and C++ because of the development of UNIX (which is written in C) versions of the software. Most modern packages also include an appropriate high-level-language programming module which can be used to customise areas of the application software to particular requirements.

Applications software CAD and CAM software applications in manufacturing fall into four broad groups.

Mechanical engineering
(1) Two- and three-dimensional mechanical computer-aided draughting (MCAD);
(2) three-dimensional computer-aided design for geometric modelling;
(3) CAE for the engineering analysis of designs; and
(4) CAM for mechanical components—mainly parts machining and fabrication.

There is now a strong movement to integrate mechanical CAD and CAM in the form of CADCAM in order to reduce errors, costs and the time taken to bring a new product to the market.

Electrical/electronic engineering
(1) electronic CAD (ECAD) for printed circuit board (PCB) and printed wiring assembly (PWA) design;
(2) ECAD for integrated and hybrid circuit design and engineering;
(3) CAE for circuit and product engineering analysis; and
(4) CAM for PCB and PWA manufacture.

Electrical CAD and CAM have been more integrated than mechanical CADCAM for some time, because their joint development was needed to enable the efficient design and production of integrated circuits and complex multilayer PCBs and PWAs.

14.4.3 Mechanical CAD and CAE

The usual application areas for mechanical CAD and CAE software are:

(1) two- and three-dimensional computer-aided draughting (CAD),
(2) geometric modelling of designs (CAD), and
(3) engineering analysis of designs (CAE).

14.4.3.1 Computer-aided draughting

CAD has the following advantages over manual draughting.

(1) Once proficient, a CAD user's overall productivity can be 3–5 times better than manual draughting. Improved advantage ratios can be obtained and are often claimed by vendors, but the lower figures give a truer average picture over all operations.
(2) Geometric shapes and features can be easily and rapidly created, copied, moved, amended and projected.
(3) Designs, shapes and features can be saved for future use on other drawings.
(4) Additional views can be created automatically.
(5) Dimensions are calculated automatically by computing the X–Y coordinates.
(6) Features such as hatching, chamfering, fillets, etc., are automatic once defined.
(7) A system of drawing levels allows segregation of assemblies, subassemblies and component details as on manual drawings, but the features can be transferred/copied between layers.
(8) The use of different colours for lines, features and dimensions improves the clarity of complicated views both on the screen and in the plotted drawing.
(9) The quality of the drawings from the plotters is usually superior to other methods of reproduction.
(10) Drawings can be produced at any scale from the same digital information.

CAD has the following disadvantages versus manual draughting.

(1) Capital investment is higher. A minimum of CAD software, a workstation and a plotter is needed to replace a drawing board.
(2) CAD uses computers, plotters, etc., that can malfunction or breakdown. Therefore, the additional costs of hardware and software maintenance agreements are necessary if lost time is to be kept to a minimum.
(3) If there is a fault with a shared piece of equipment, e.g. a fileserver or plotter, then several people will be inconvenienced.
(4) Training requirements vary with the complexity and user-friendliness of the software, but several days are usually needed to cover training on the most frequently used features. Even with this it will normally be several months (depending on the usage rate) before a designer becomes fast and efficient in using the program.
(5) It takes several minutes to bring a system on line and to produce a required drawing. This cannot compete on speed and convenience with manual drawing when the requirement is for a small and simple sketch. N.B. three-dimensional sketching packages that are very versatile are just coming on to the market, but these are additional to the basic two- and three-dimensional packages and cost approximately £800.

The operation of CAD systems In CAD the graphics display screen replaces the two-dimensional drawing board. All programs provide a basic two-dimensional draughting package which varies in the functionality offered. Some programs also offer three-dimensional draughting by adding depth and perspective to the two-dimensional base by various methods.

CAD programs incorporate a large number of features that enable the creation of the geometric shapes, symbols, dimen-

sioning systems and informational elements normally used in design and drawing. In addition, they offer drawing operations and options that derive from computer graphics technology—these include moving drawing elements around the screen, choice of many colours for any line or element, choice of line widths and style, different text sizes and fonts, view zooming, and other image-handling operations.

CAD systems are operated via a system in which the available features are presented in listings and sublistings called 'menus'. The desired feature is selected by placing an arrow or cursor cross-wires over the name of the feature in the menu which is displayed on the screen or a graphics digitising tablet. The feature selection and drawing actions are usually input by the user via either a stylus or puck on a digitising tablet, a mouse, tracker, ball, joystick or light pen. The computer keyboard is, by comparison, used only occasionally for the input of dimensions, data or text. More details of the hardware mentioned are given in Section 14.4.6.

Some basic menu elements are shown in *Figure 14.13* and a typical menu screen is shown in *Figure 14.14*.

Main menu	*Submenu*	*Purpose*
FILE	CONTROLS GENERAL MANAGEMENT OF FILES	
	help	Provides guidance on all features
	status	Indicates status of the system
	directory	Directory of files
	create	Open a new drawing file
	load	Load an existing drawing file
	return	Return to prior drawing
	save	Save the working file
	save as	Save the working file under another name
	copy	Copy the current drawing
	rename	Rename the current drawing
	end	End the working session
VIEW	CONTROLS VIEWING OPTIONS OF THE SCREEN	
	2D	Two-dimensional perspective working
	3D	Three-dimensional perspective working
	zoom	Zoom view in and out
	pan	Move view around
	redraw	Redraw the current screen
GEOMETRY	PROVIDES THE DRAWING GEOMETRY FEATURES	
	lines types	Various types of line can be drawn
	arcs methods	Various methods of production
	circles methods	Various methods of production
	ellipse methods	Various methods of production
	polygon methods	Various methods of production
	splines methods	Various methods of production
	tangent methods	Various methods of production
DRAW	PROVIDES THE ACTIONS USED IN DRAWING	
	trim	
	snap	
	rotate	
	mirror	
	stretch	
	fillet	
	chamfer	
EDIT	EDITS DRAWING OR FEATURES	
	erase	
	move	
	copy	
	undo	
	re-do	
ATTRIBUTES	PROVIDES PERIPHERAL FEATURES TO DRAWING	
	levels	
	colours	
	text	
DIMENSIONS	CONTROLS TYPE AND STYLE OF DIMENSIONING	
	inch	
	metric	
	+/−	
	geometric	
	arrows	
OUTPUT	CONTROLS METHODS OF OUTPUT	
	plot	
	print	
	export	

Figure 14.13 Basic features included in most CAD draughting packages. Most CAD systems contain basic features which are expanded as the specification increases on more expensive packages. Features are accessed by the selection of the feature name from a menu listing displayed on the screen. The menus have a hierarchical structure with top-level menus splitting down into submenus and sub-submenus. The nomenclature and range of features varies between software packages

14.4.3.2 Geometric modelling

Geometric modelling is used to provide three-dimensional constructions which:

(1) produce a realistic three-dimensional image of the design to give better visualisation;
(2) provide data that can be used in calculating various engineering analyses of the design;
(3) provide data for the subsequent production of detail drawings; and
(4) provide output for DNC, CNC and other CAM processes.

Modelling can give tremendous advantages in the design process because it can save expensive and time-consuming experimentation and the reworking of unsuccessful designs. Three methods of producing three-dimensional models are used: wireframe, surface and solids.

Wireframe modelling This method creates a model by connecting points in space. The connecting lines so drawn are perceived by the viewer as being the edges of the viewed object. Initially all lines are shown. This gives an effect of transparency in which features such as back edges that would normally be hidden from the viewer can be seen. Such an image is acceptable with simple objects but can become confusing with complicated ones. However, a computer graphics technique of hidden-line removal is employed which clarifies the image and seems to present only the normally seen surfaces. This is achieved by the viewer's visual interpretation of the graphic, not a mathematical construction of the model. Most CAD packages can now provide automatic surfacing of wire frames which can be rendered in colour shades to produce images with surfaces, perspective and depth. A wireframe model is shown in *Figure 14.15*.

Surface modelling This involves the construction of models by the mathematical definition of surfaces. Surfaces are defined by applying various 'surface element' functions such as: surface of revolution, tabulated cylinder, fillet surface, or free-form sculptured surface. Surface modelling is particularly useful in modelling complex sculptured surfaces such as aircraft sections, car bodies and mould forms for plastics and castings. The surface data can be taken for use in CAM machining modules for the production of part program tool paths in either DNC or CNC operations. This is particularly useful in the multi-axis machines that are needed to create complex surface contours.

Solids modelling Two methods are used: constructive solid geometry (CSG) and boundary represention (B-Rep).

CSG combines simple geometric elements to form more complex shapes. The elements are known as 'primitives' or 'entities' and are in the form of blocks, cones, cylinders,

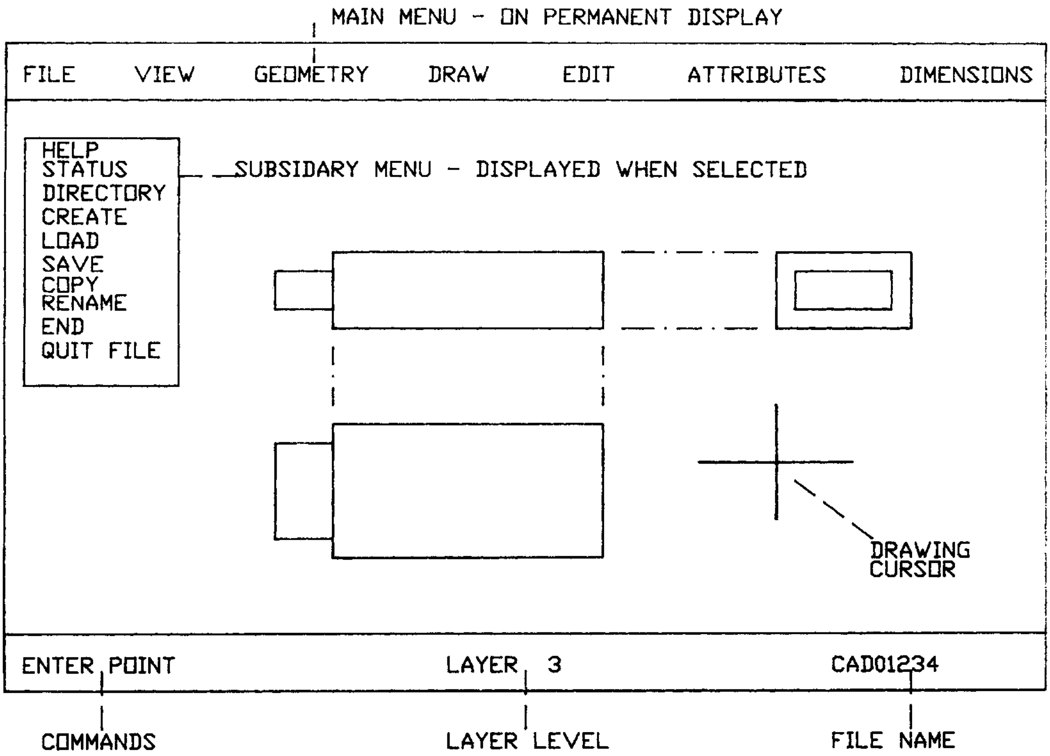


Figure 14.14 Layout of a CAD screen

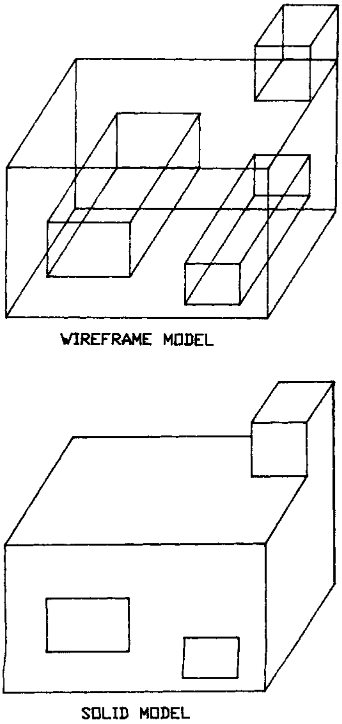


Figure 14.15 Types of model construction

spheres and tori. These can be combined in the Boolean relationships of union, difference and intersection. This method works better with regular shapes than with irregular shapes. The Boolean combinations are illustrated in *Figure 14.16*.

B-Rep uses boundary data, i.e. definitions of the constituent surfaces/planes, to define the solid model. The program is designed to define how the individual details are joined together. Complex non-uniform shapes need extensive

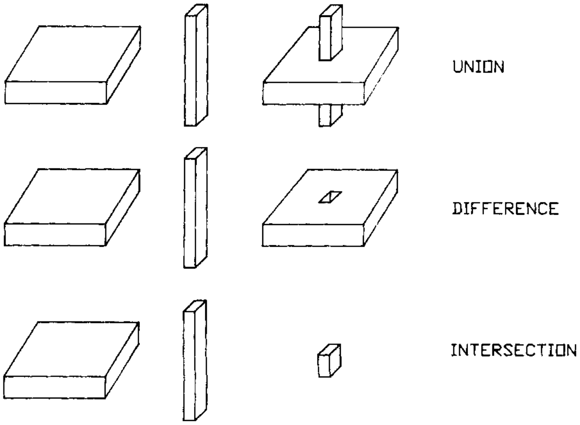


Figure 14.16 Boolean operations

description of the boundaries of the faces, but others with uniform cross-section can be more easily created by developing the shape from the appropriate axes (e.g. with solids of revolution such as cylinder, the radial cross-section can be swept for 360° around the centre line). Uniform bodies may be 'extruded' to any desired length.

Solid models can also be created from conversions of wireframe models, surface models and two-dimensional details.

14.4.3.3 Computer-aided engineering

Computer-aided engineering (CAE) encompasses many applications in the engineering analysis of designs. This work originally needed mini or mainframe computers because of the size of the programs which contain many calculations, but with the increased capacity of workstations and PCs many CAD suppliers are now including CAE modules in their program suites. This is possible because the data used to construct the three-dimensional models can also be used for a variety of engineering analyses, e.g.

(1) mass properties—volume, weight, centriods and moments of inertia;
(2) stresses;
(3) thermal properties;
(4) interferences; and
(5) movement paths.

The major tool for carrying out these analyses is finite element analysis (FEA). This is a technique that breaks the model down into small simple elements of a size appropriate to the local calculation of the condition being analysed. Each element is considered to be joined to its neighbour at points called 'nodes', and may be one, two or three dimensional. The network of elements over the whole model is known as the 'finite element mesh' (FEM) (see *Figure 14.17*).

FEA involves a considerable amount of work in the model definition and data preparation before the actual element analysis takes place. Interpretation of the results of the analysis also creates a large workload. Consequently, software houses have developed pre-processors and post-processors to handle the preparation and interpretation tasks. These may be used in conjunction with a resident FEA program from the software supplier or be interfaced to other proprietary FEA packages.

14.4.3.4 Future trends in CAD and CAE

Software will continue to show the normal development trend of the lower specification programs gradually being upgraded to incorporate features currently available only in higher specification programs. Relatively cheap CAD software suites will incorporate more parametric design, geometric modelling, full-colour three-dimensional visualisation, engineering analysis and kinetic simulation modules.

Two-dimensional draughting has until now been the most common application, but three-dimensional draughting and modelling suites with engineering analysis options will become far more widespread. Three-dimensional sketching packages with links to databases and spreadsheets will come into use for initial design work.

Open systems will become widespread with all hardware and software products being tailored, where necessary, for open working.

The majority of applications programs will be written with graphical user interfaces (GUIs) of the MS-Windows (TM) type of cascading menus.

High-specification PCs will run on UNIX or a variant of it as well as on DOS, and UNIX will be consolidated as the most widely used operating system.

Low-cost workstations and high-specification PCs will compete on price and be the major platforms for CAD and CAM systems. These will provide reduced instruction set computers (RISC) or 80486 and later microprocessors, increased operating speeds, larger memories and enhanced graphics. 64-bit processors will be put into machines in this size range. The distinction between workstations and high-specification PCs will disappear.

14.4.4 Computer-aided manufacture

14.4.4.1 Introduction

In the widest sense CAM can be taken to encompass all CAM technology including numerical engineering, flexible manufac-

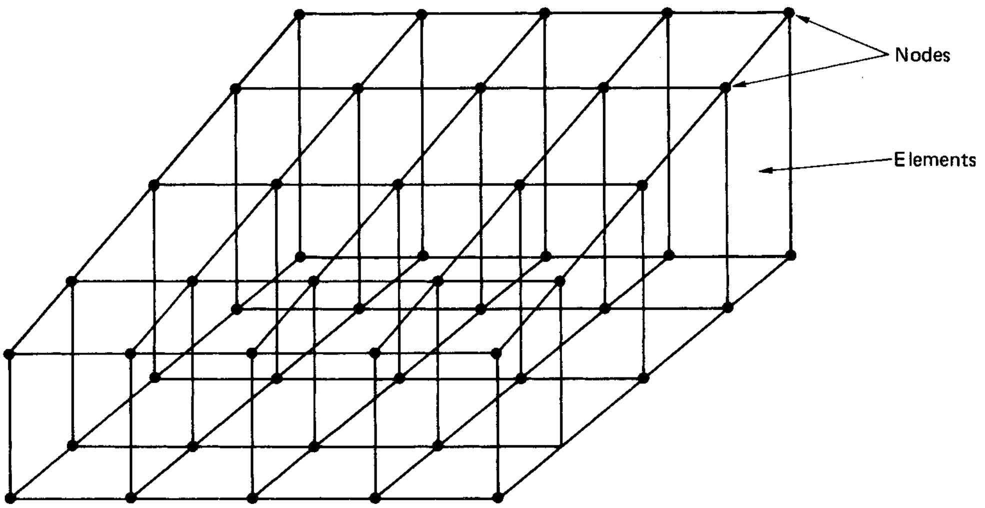


Figure 14.17 Finite element mesh

turing systems (FMS), flexible assembly and robotic welding and handling. The narrower and more usually understood scope of CAM within the context of CAD/CAM is in the computer-aided programming and related engineering activities for numerically controlled manufacturing processes. The narrower definition is used here because FMS and robotics are discussed in other sections.

CAM packages are offered to the market in three forms: CAM, CAMCAD and CADCAM packages.

CAM packages are entirely separate from any CAD packages. In this case a series of instructions is entered into the computer to define the geometry of the part. This early technique has been largley overtaken by the use of interactive CAD facilities embodied in CAMCAD and CADCAM packages.

CAMCAD packages contain an in-built CAD program which is used to define the part geometry. This is done by the engineer working with a graphics screen to draw the part in a similar fashion to CAD draughting. In more recent packages, these facilities have been greatly enhanced to the point where some three-dimensional work is possible. Modern packages also include the communications capabilities (IGES and DFX, etc.) to import part geometry from other CAD systems. These programs are known as 'CAMCAD' packages because they have been developed from early CAM programs by companies primarily serving the numerical control market, not the product design market. The structure of a CAMCAD package is shown in *Figure 14.18*.

CADCAM packages contain CAM facilities which comprise a series of modules that are integrated with the CAD within a complete CADCAM suite. In this case the part geometry is created in the CAD suite and held in a common database from which it can be downloaded to the CAM modules. This is a more efficient and potentially more error-free method of working than using separate CAD and CAM.

In each of the above cases the type of numerical control machine requires specifically tailored software. This is supplied in the form of modules that can be purchased separately, usually as additions to a mandatory base module. The most commonly used modules are:

(1) milling—($2\frac{1}{2}$-dimensional and three-dimensional, surfaces, and multiplane modules);
(2) turning (two axis, multi-axis and C axis);
(3) grinding;
(4) sheet metal punching;
(5) laser cutting; and
(6) electrical discharge machining.

Typical screens for milling, turning and sheet metal work are shown in *Figures 14.19* to *14.21*, respectively.

14.4.4.2 Software

Part programming has been computer aided since the introduction of the first automatic programming tool (APT) languages, but those early programs did not have the advantages of the more recently developed interactive graphical programming. This technique is now becoming the standard in industry with the development of powerful workstations and PCs that include good quality colour graphic screens.

Interactive graphical programming provides immediate visual interaction between the programmer and the computer. Instructions are input via the selection of commands from command menus displayed on the screen or on a digital tablet. The results are shown on the graphics screen in the form of drawings, geometric elements, dimensions and text displays, giving the programmer an immediate two- or three-dimensional indication of the action just taken.

The major elements of CAM software are:

(1) part geometry definition;
(2) tool definition, libraries and management;
(3) production of the cutter location file (CLF);
(4) post-processing of the CLF;
(5) tool-path verification displayed on screen;
(6) storage of parameters for use on other parts;
(7) DNC and other communications links; and
(8) production control facilities.

Part geometry definition If the part geometry cannot be imported from a CAD file it must be created by the programmer. To this end, the CAM software contains CAD features that allow the production of part shapes and automatically registers the commanded and resultant dimensions and coordinates.

Tool definition The cutting tools and their dimensions must be defined in order to allow the correct calculation of tool paths.

Cutter location file When the part geometry and tooling data are completed, the software writes the program which defines the cutter path. This is usually known as the 'cutter location file' (CLF).

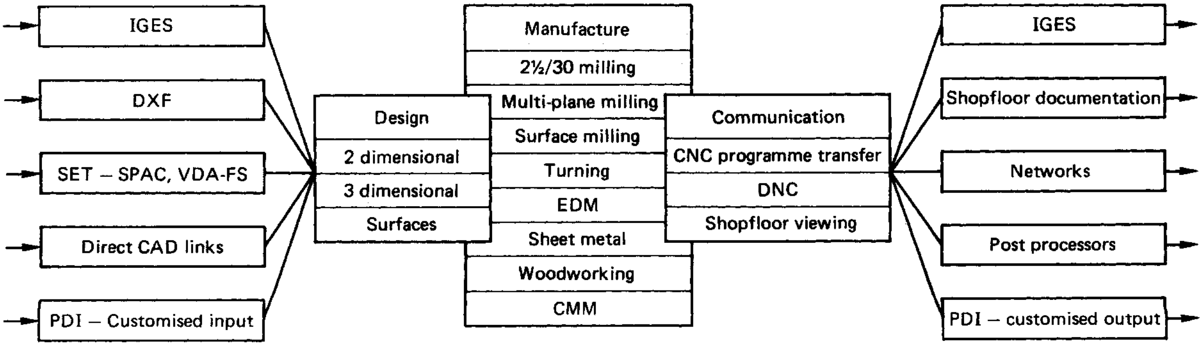


Figure 14.18 CAMCAD system. (Courtesy of Pathtrace Engineering Systems)

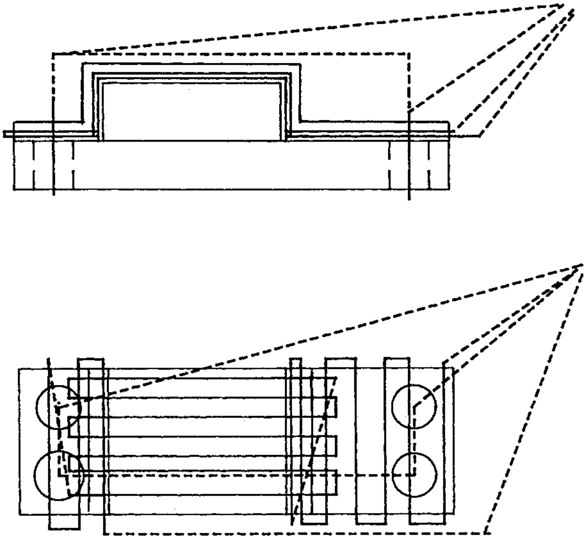
Post-processing The computer produces a CLF that must be converted to the input requirement of the specific machine to be used. The conversion operation is carried out by a separate software program known as a 'post-processor'. There are two ways in which this task is handled.

(1) Post-processors that are specific to a particular machine and must be purchased (normally from the machine supplier) for each machine type. These are the type used in smaller and cheaper CAM installations.
(2) A universal post-processor that can be customised to provide a post-processor for the machine to be used. These are normally available with advanced CAMCAD and CADCAM packages.

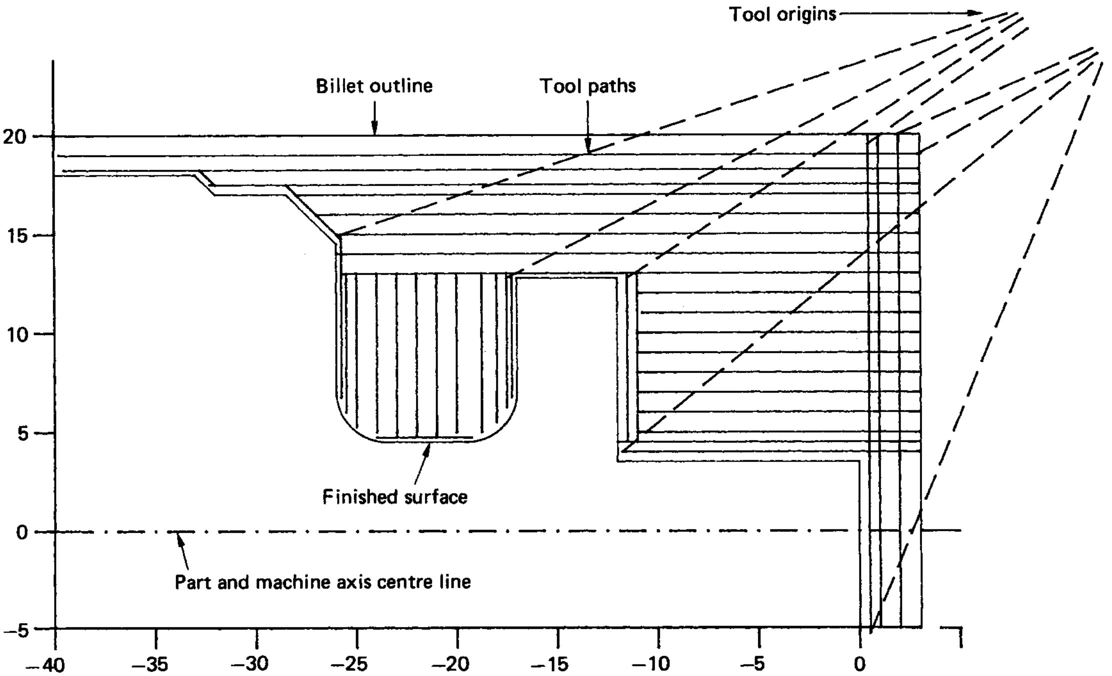
Tool verification Once the tool paths have been calculated, they can be illustrated on the screen, in sequence, so that the progress of the machining operation can be seen. Verification is available in both two- and three-dimensional solids presentations. Automatic verification will check for tool collisions with the work piece, clamping, fixtures and machine parts.

Storage of parameters for other parts It is possible and very useful to have the parameters of a part available for future use in the programming of a similar part or family of parts.

DNC and other communications links Complete programs may be down-loaded at one transfer if the machine tool buffer memory is large enough. If not (normally only the case with large programs), the blocks can be drip-fed to the machine. In earlier systems, links between the host computer and machines was created by the addition of a receiving unit on each machine and a manual switching box to route programs to the correct machine. The current trend is to add a communication card to the machine control and route the programs via a fileserver which controls a LAN running to all the machines.



(a)



(b)

Figure 14.19 (a) Screen for 2½-dimensional milling and drilling operations. (b) CAM screen for turning

(a) (b)

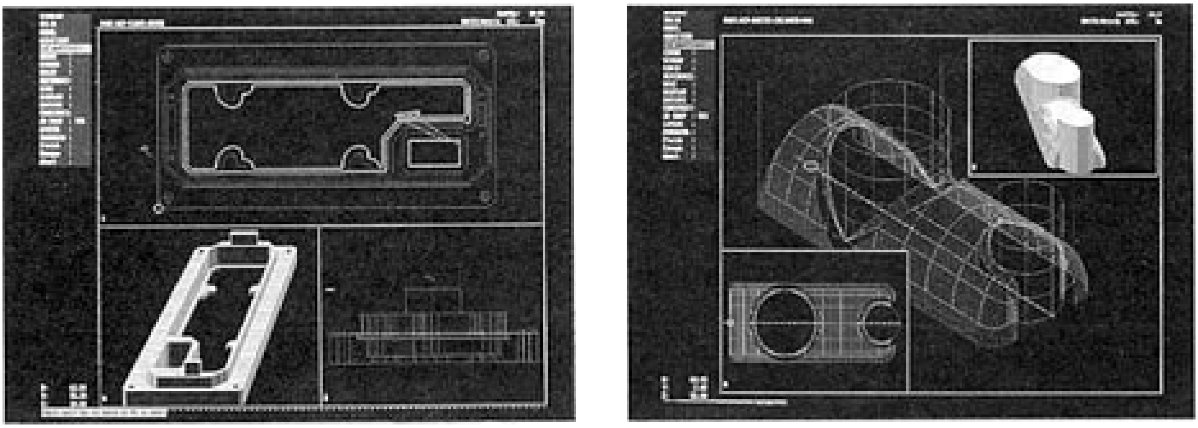


Figure 14.20 (a) Screen for 2½-dimensional milling. (b) Screen for surface milling. (Courtesy of Pathtrace Engineering Systems)

Production control facilities These provide the capability to control and gain data on production batches on the machines that are linked to the workstation or host computer by a DNC link or LAN. The trend to integrate production control activities with the programming work will increase as CIM becomes more widely implemented.

14.4.4.3 Hardware

CAM systems hardware is very similar to CAD hardware with the addition of tape punching. In an integrated CADCAM or CAMCAD system all the activities are usually performed on the same equipment. A typical system comprises the elements described below. For further details of CAD and CAM hardware see Section 14.4.6.

These should comprise the following elements.

Workstations
(1) Central system—with a processor of at least 80286 but preferably 80386 level. Also a maths coprocessor should be included to increase the speed of the many calculations involved.
(2) Graphics screen—(needed for part and tool path representation in interactive working with the programmer). A video graphics adapter (VGA) grade should be used with a resolution of at least 1024 × 768 for continuous use.
(3) RAM—should be a minimum of 640kbyte with 1Mbyte preferred for heavy use. (1 or 2Mbyte are now becoming standard offerings).
(4) Keyboard—either standard (83/84) or enhanced (110/111) types.
(5) Input device—either a digital tablet, mouse, tracker ball or similar device for part geometry definition. Digital tablets or a mouse are becoming the preferred method due to ease of handling.

Tape punch This is needed to produce new and duplicate numerical control tapes.

Plotter This is used to plot product outlines and tool paths. It is a very useful visual aid. It can also be used for associated jig and tool design.

Printer This is used to print copies of the numerical control programs, tooling and other data listings.

Tape streamer This is used to take back-up copies of files. Back-up can be done on floppy or hard discs, but a large system will need a tape streamer.

File server This is needed in a network of several workstations and peripherals and for communication to the machines in a DNC environment.

Communications Serial communications ports for RS232C, RS422 or Ethernet or token ring LANs should be provided for linking to numerical control machines. A parallel port should be used for connection to a printer. Communication methods have limitations on the distance between elements of hardware and the total network length. These need to be checked against the total and likely future requirements and cost limitations.

14.4.5 Electronic CADCAM

14.4.5.1 Introduction

The rapid expansion of the electronics industry resulted in the early development of electronic CAD and CAM for the integrated design and manufacture of electronic products. This is because the combined effect of micro-sized and multiple-layered circuits has made the use of computer-aided techniques essential for efficient and high-quality design and production, as the tasks became too complex and miniature to be either possible or economical to do by manual methods.

The software is completely different from mechanical CADCAM software, but the hardware used is generally the same, requiring an equivalent quality of graphics screens, RAM capacities and mathematical calculation handling via maths coprocessors. For a description of CAD hardware see Section 14.4.3.

A large range of computer-aided applications is available, one of the most widely used being the design and production

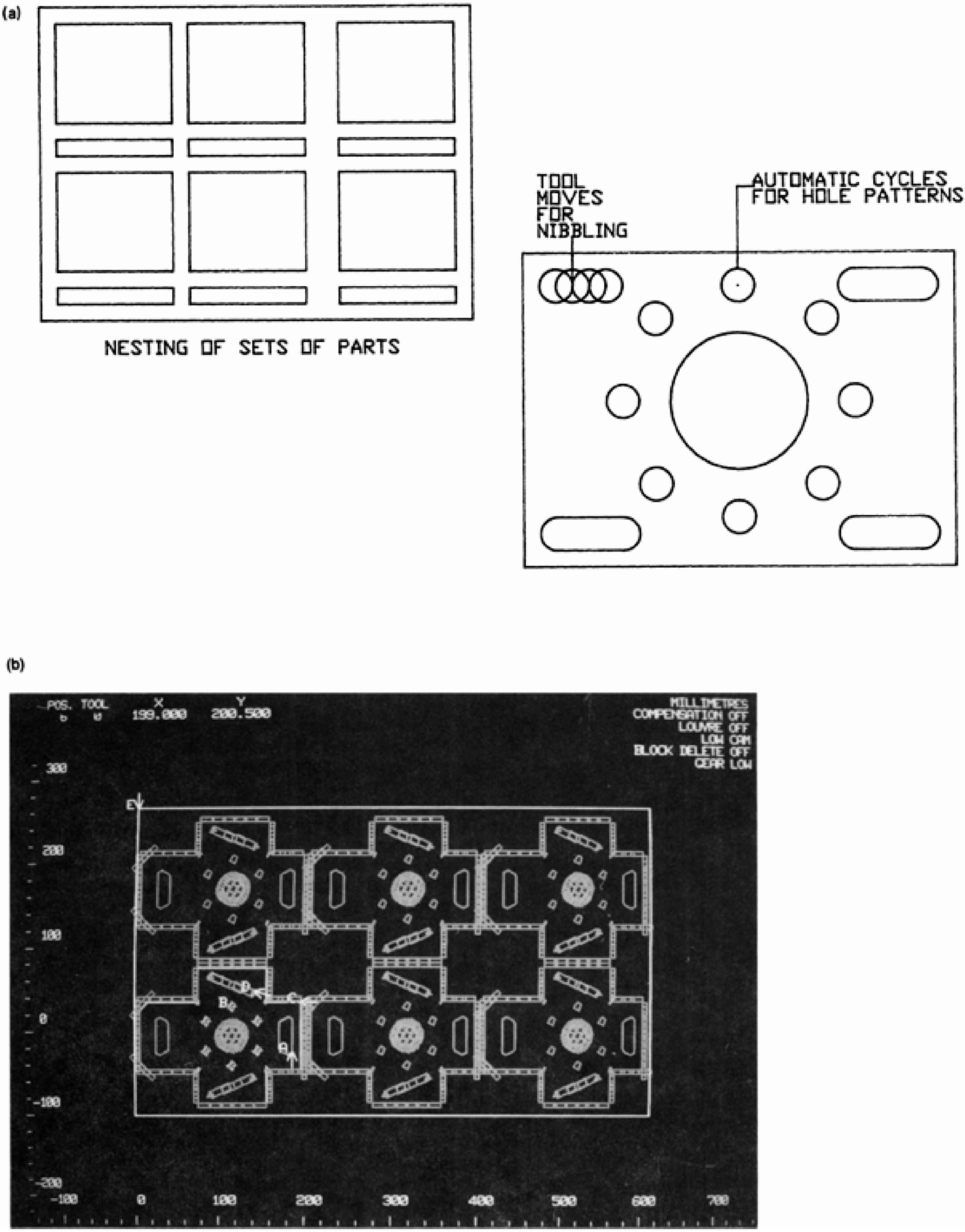


Figure 14.21 (a) Screens for sheet metal operations. (b) Screen showing sheet metal punching. (Courtesy of Pathtrace Engineering Systems)

of printed circuit boards (PCBs). As with mechanical CADCAM, the capabilities of electronic CADCAM programs vary with price, the more expensive packages (£3000 and above) offering a greater degree of automation, flexibility and analysis than basic programs costing a few hundred pounds.

14.4.5.2 CADCAM systems for PCB design and production

The main elements of a CADCAM system for PCB design and production are described below.

Schematic design capture This is an interactive technique in which the designer constructs the proposed circuit by placing component symbols and connections on the screen. This is a functional representation of the circuit design intent, not a physical layout. The great advantage of the technique is that it allows the use of predrawn symbols, thus saving considerable draughting time, and facilitates easy editing and modification of the design via normal CAD methods adopted for this special application. Schematic capture programs can also contain error-checking facilities which check for errors such as wires going nowhere, shorted outputs and bus contention.

Component/symbol database Large system databases can hold details of over 2000 components and 100 pad shapes which are reproduced on the schematic design when selected by the designer. The data held include component name, type, value, and pin number. Such a database is usually organised as a number of sublibraries, one for each of the common types of component used. A variety of coding systems is used but the best employ the industry standard numbering systems that define various component types and ranges, e.g. the 74000 series of components.

Net lists These provide hard copy of data about the circuit and are produced automatically from the calls on the component database made when producing the schematic. Commonly used lists include bills of materials, connection tracks and pins.

Circuit layout The schematic is converted into a physical layout of components on the board. The components can be placed either interactively by the designer or automatically.

Routing This defines the routes of the tracks between the components and around the board. This may be done interactively or automatically. Some routing programs use a grid as the basic template for defining routes, whilst others work without a grid by using special algorithms. Algorithms can give denser track layouts because they are not restricted to the pitch of the grid.

Outputs from the design process Only a computer can efficiently and accurately produce, modify, align and keep track of the coordinates and geometry of the successive layers of the board.

(1) *Output to photo-plotters*—Photographic/chemical processes are widely used in the manufacturing of a number of electronic products. In PCB production these are required for producing the complex layouts of the copper tracks, connection holes and solder resist layers. The most popular method of converting the design data to a format suitable for use by a photo-plotter is by converting the net lists to produce Gerber format files. these are used by the PCB manufacturer (often a specialist subcontractor) to produce photographic negatives which become the master artwork for the production of the various layers of the PCB fabrication. Gerber editors and viewers allow the lines and holes that are to be photo-plotted to be seen prior to production.

(2) *Output to NC drilling machines*—The holes needed to connect tracks of copper from one board layer to another are normally produced by special NC drilling machines. The coordinate positions of the holes are produced from the design process and transmitted as a file or tape to the NC machine control.

(3) *Output of drawings and documentation*—Drawings of the various stages of the design process and for each layer are plotted from the CAD system. Bills of materials, and net and other listings are printed.

Engineering analysis Programs are available for the analysis of various conditions of the circuit and PCB designs. These include thermal stress, impedance and signal noise analysis.

Circuit simulation Simulation programs are available for analogue, digital and mixed circuits. They can save considerable time and cost in experimentation and bread-boarding by creating the circuit operation and answering 'what if' questions. Simulation programs use net lists from the schematic and device data from libraries as inputs and apply mathematical modelling by the use of algorithms from a library of algorithms.

14.4.6 Hardware review

The range of specifications, sizes and technologies employed in CAD and CAM hardware is enormous. Consequently this section can only give an indication of the more commonly used hardware and typical specifications.

14.4.6.1 Workstations/host computers

System workstations and host computers fall into three categories:

(1) PCs—general-purpose PC designs adapted for CAD and CAM work (60% of installations and increasing);
(2) workstations—purpose designed for CAD and CAE (20% of installations);
(3) mini computers for large and high specification systems (19% of installations).

PC hardware Early PCs were mainly used for stand-alone or small networked systems using low or medium specification CAD and CAM software, whilst higher specification programs were run on workstations, mini computers or mainframes. With the introduction of 16-bit processors and larger memories, high specification programs can now be run on PCs.

The major PC platforms for CAD and CAM work are now machines containing Intel 80386 and 80486 or equivalent processors. Earlier speeds for these were 12 and 16 MHz, but 20, 25, and 33 MHz are available and better suited to CAD work. An accompanying chip for maths coprocessing (e.g. Intel 80287 or 80387) is also needed because of the high number of calculations involved in graphics work.

Previously, systems were supplied as standard with up to 256 kbyte system RAM with extensions to 640 kbyte. This is adequate for simple applications but 1 Mbyte system RAM or more is advisable when handling sophisticated applications. Manufacturers now offer 1, 2 or 4 Mbyte as standard with optional expansion capability to 8, 16, 32 and 64 Mbyte of RAM.

The minimum amount of video RAM needed for the currently available editions of software is 256 kbyte and this should suit most two-dimensional applications. 512 kbyte and

1.2 Mbyte of RAM is becoming used as additional features are introduced. Three-dimensional working and engineering analysis modules need a large memory capacity. Video RAM is housed on graphics adaptor cards of which there are many models and suppliers. It is possible, therefore, to upgrade graphics capabilities without replacing the central system. Some graphics cards have expansion capabilities on board.

Hard disc storage is the most frequently used medium for mass storage in PC-based systems. The minimum storage for a simple system without a heavy load is now probably 40 Mbyte, and the price of disc storage is now so reduced that a purchaser will probably be offered this as standard with options for 60 and 80 Mbyte. Fileservers addressing several machines should have between 200 and 1000 Mbyte (1 Gbyte), depending on the size of the system and its load.

The performance of systems is greatly affected by the memory access time and bandwidth. This is particularly so in 32-bit microprocessors with high clock speeds where the processor will continually be waiting for access to memory. To avoid this condition, additional high-speed (cache) memory is used as a buffer between the processor and memory. The cache memory may be built into the processor chip or on the main system board.

Due to the large number of IBM or IBM clones installed, the most commonly used operating systems are versions of disc operating system (DOS) which was originally written by Microsoft for the IBM PC—the IBM version is called 'PC-DOS'. This and other PC-oriented systems were designed to control single-computer installations running one task at a time and are not suitable for controlling multitask or multiuser systems. Modern PCs are, however, providing an in-built migration path from DOS systems to multitask operating systems such as OS/2 and miltuser operating systems such as Novell and UNIX.

The more powerful UNIX based PCs are challenging the use of entry level and bottom of the range workstations as the specifications of each become less distinguishable.

Workstations Technical workstations have been specifically designed to incorporate the enhanced processing power, memory, graphics monitors and input devices needed by medium- and high-specification CAD and CAM systems. Workstations normally use a multitask operating system such as UNIX, XENIX or VMS (DEC systems).

Workstations can be used as stand-alone units, connected to a plotter or printer. More usually they are incorporated into larger systems where they may be networked to common output and storage peripherals or act as a satellite terminal to a central system computer.

In the past, workstations were sufficiently expensive (approximately £60 000 to £100 000) to deter small and medium sized companies from investing in a higher level of CAD and CAM. However, hardware prices are reducing, particularly with the introduction of reduced instruction set computer (RISC) architectures. There is a very strong trend in this direction being pursued by many workstation manufacturers and the general availability of less expensive workstations (at less than £3000) that are indistinguishable from top-specification PCs will increase greatly.

Mini and mainframe computers In the past, large computers were required to provide the necessary processing, memory and mass storage capabilities needed for medium and large installations running sophisticated software. The need for a mainframe computer restricted applications to large or high-technology companies in industries such as automobiles and aerospace. The introduction of mini-computers widened the user base substantially, with systems such as the VAX series from DEC. With the increasing use of workstations and PCs as system processors rather than satellite terminals, the need for mini and mainframe computers is now reduced to all but the largest systems.

14.4.6.2 Graphics display monitors

The graphics monitor is the major vehicle for interaction between the designer and the CAD program. In addition to being a substitute for the two-dimensional drawing board it is also a means of presenting three-dimensional images and the command menus used to operate the system. The sizes normally used are 14, 17 and 20 in. for low- and medium-specification systems. High-specification systems may use 26 in. displays.

Two types of monitor in common use are *raster* and *vector* displays. Both use the principle employed in television sets; a cathode ray tube projects a beam of electrons onto the internal surface of the screen, which has a coating of phosphor. The electron beam causes the phosphor to glow and it is this phosphorescence that the viewer sees. The effect lasts for only a fraction of a second and so, in order to maintain a seemingly constant image, the whole screen must be 'refreshed' with further passes of the beam at a rate of 30–80 kHz in the horizontal and 30–120 Hz in the vertical, depending on the model.

Raster displays These are the most commonly used type for commercial CAD and CAM displays. The majority of CAD-CAM displays are colour monitors in which separate phosphor dots emit red, blue or green light in response to electron excitation. The display contains lines of *pixels* which contain groups of three points of phosphor arranged in a triangle (triad). The pitch of the triads (known as the 'dot pitch') is typically 0.28 or 0.31 mm. By digitally directing three beams (one for each colour) at the pixels and by varying beam intensities, it is possible to create a wide range of colours based on mixtures of the primary red, blue and green. Early models could only display four colous, but 16, 64 and 256 colours are now common on PCs and 4096 colours on higher specification machines. The number of colours depends on the number and processing methods of the bits used per pixel, not on the screen specification.

The *resolution* of the display monitor is a measure of its clarity, which is dependent upon the number of pixels in each horizontal line and in the vertical. Resolution is expressed as the number of pixels and is a function of the number of bits used to program the pixel, not the quality of the screen. If 4 bits are used per pixel, then a display with a resolution of 640 × 480 pixels needs:

$$640 \times 480 \times 4 = 1\,228\,800 \text{ bits}$$

at 8 bit/byte

$$= 153\,600 \text{ or } 150 \text{ kbyte of video RAM}$$

The development of improved resolution with successive models of monitor is shown in *Table 14.1.* Currently VGA is the normal standard supplied with super (SVGA) and extended super (ESVGA) gaining increasing usage. The earlier models are now practically unobtainable.

The overall performance of the graphics display is dependent on the processing power (now housed in one chip for a VGA display) and the amount of random access memory (RAM)—typically 256 kbyte to 1.0 Mbyte is used. These may be housed on the main system board or on a separate graphics adaptor card. There are many models of such cards on the market and these can be purchased to enhance the performance of the graphics.

Table 14.1 Resolution of graphics monitors (graphics and colour mode, not mono and text)

Monitor type	*Maximum resolution*	*No. of colours*
PC colour graphics adaptor (CGA)	320 × 200	4
PC enhanced graphics adaptor (EGA)	640 × 350	16
PC video graphics array (VGA)	640 × 480	16
PC super video graphics array (SVGA)	800 × 600	256
PC enhanced video graphics array (EVGA)	1024 × 768	256
Workstations	1280 × 1024	≥1000

Vector displays These were developed before raster displays and were first applied to the early monochrome screens. In this type of monitor, a single electron beam draws a continuous straight line between two coordinate points anywhere on the screen. With a typical resolution of 4096 × 4096 the line definition is better than in the raster type because vector displays are drawn with a single sweep of the electron gun, whereas the raster method must employ a group of pixels. The reduced resolution on raster displays shows up particularly on diagonal lines the edges of which are in the form of a series of steps.

14.4.6.3 Input devices

Hand held interactive devices The main method of communicating with the computer for CAD and CAM work is the use of various hand-held devices whose input produces an immediate response on the graphics screen. This allows the designer to work 'interactively' with the computer.

These devices control the position of a cursor arrow and/or coordinate cross-wires to produce the desired geometry 'drawing' and other images on the screen. They are also used to select the program menu options by pointing the cursor/cross-wires at the desired option description from a list (i.e. menu) displayed on the screen. There are several types of device in general use.

Digitising tablets are flat boards which represent the drawing and screen area. Any coordinate point within a defined area can be selected and transmitted to the computer. Digitising tablets are normally used where resolutions of better than 0.015 in. are needed for the transfer of existing drawings (i.e. digitising) into the computer. Several methods of tablet construction and point definition are available, but the most widely used has a network of fine wires beneath the surface of the board and a hand held stylus or puck. When these are placed over a point, its *x–y* coordinate position is defined by the electrical interaction of the wire mesh and the hand-held device. Pressing the end of the stylus on the tablet activates a switch. A puck has cross-wires which are placed over the desired point. Keys are used for control, as in a mouse.

Floor mounted A0 size digitising tables are available, but the usual size for CAD and CAM is a desk-top tablet of between approximately 12 in. × 12 in. to 15 in. × 18 in. Tablets normally have an area dedicated to the drawing activity and another area displaying command menus and possibly a full alpha-numeric key layout. This is indicated on a printed template and the stylus or puck is placed over the printed command or key in order to select it.

Tablets are often part of the larger work surface found in specially designed CAD workstations.

A *mouse* is a very widely used device, particularly on PC-based systems. A ball or set of wheels is housed in a small hand-held device. As the device is moved across any flat surface, the ball or wheels rotate through friction with the surface. This rotation produces signals that are converted to indicate the movement of the mouse relative to its starting position. Function keys incorporated in the mouse are then used to give 'enter' or other commands after the movement (which is displayed instantaneously on the screen) has been completed or menu selection made. Early models were analogue devices using variable-resistor systems. Most modern devices are digital with resolutions of up to 400 dots per square inch (dpi) (0.025 in.).

Tracker balls, *thumbwheels* and *joysticks* work in the opposite mode to a mouse, in that a base unit remains static and a large roller in the tracker ball, thumb wheels or joystick are rotated in the base by hand action. Tracker balls and joysticks produce a combined *x* and *y* movement whilst thumbwheels produce an *x* or *y* movement—the latter can be an advantage in work requiring accurate location.

Light pens are similar to a ball-point pen in shape. When the pen is held against the screen and an internal switch activated it detects the presence of light. A signal is transmitted to the processor which can then deduce the position of the pen. The accuracy of a pen is not as good as other methods which limits its use for drawing work. It is used, however, as a general pointing device. Pens are not suitable for all screens which is another factor that limits their wide use.

Keyboards Keyboards are needed to input text, numerical data and some commands but are not the major means of interaction between the user and the computer (this is normally a mouse, digitising tablet or other similar hand-controlled input device).

Two layouts of keyboard are in general use in the UK. Both use the QWERTY layout for the alphabetical keys but have different arrangements for the numeric/cursor keypad and function keys. The two layouts are:

(1) *83/84 type*—this was the first keyboard developed. It has 83 keys in the US version and 84 in the UK version (the additional key being required to accommodate the £ sign). There are 10 function keys. This board transmits 8 bits on each depression and another 8 bits upon release of a key.
(2) *110/111 type*—known as an 'enhanced keyboard' this has 110 and 111 keys in the US and UK versions, respectively. It has 12 function keys and additional control and cursor keys. It transmits 11 bits for each key depression and release.

Because of the different bit scan rates, the keyboards are not interchangeable. Either can be used with CAD and CAM systems, but the enhanced keyboard provides separate cursor and numeric key operations and more functionality.

Various suppliers assign different tasks to the function keys and operations obtained by combinations of the shift/alt/control and alpha-numeric keys. These are normally indicated by additional inscriptions on the keys and/or templates that can be laid around the edges of the keyboard.

14.4.6.4 Output devices

Plotters Plotters are used to produce line drawings from CAD or CAM system plot files onto a variety of media including drawing paper, glossy paper, polyester film, vellum and transparency film. Most plotters accept the plot file format Hewlett Packard Graphics Language (HPGL), but other formats used include Calcomp 960/970, Versatec V80 and VRF. Plotters have a memory known as a 'buffer' to hold the plot files downloaded from the computer. Standard buffer sizes range from a few kilobyte on small machines to 20 Mbyte

on top-of-the-range models. RS232C serial and Centronics parallel interfaces with the computer are normally available.

'Intelligent' plotters also have a microprocessor incorporated. This takes on the management tasks, vector-to-raster conversion and the creation of shapes such as circles, arcs, etc. from a library of 'primitives'. These capabilities release the system computer from such tasks and increase the speed of throughput. Typical resolutions for plotters are 200, 300 and 400 dpi or 0.25 and 0.0125 mm.

Machines vary in size from standard A4 up to standard A0 with larger machine sizes up to 2 m × 3 m. A4 and A3 models are usually desk mounted and the larger machines floor mounted.

Several types of plotter are in use. Pen plotters are the cheapest device for producing coloured line drawings because they only have to carry different coloured pens to do so. Alternative designs such as electrostatic, thermal and laser plotters are better for producing areas of colour shading because they use a raster technique similar to the graphics monitor principle. This necessitates the inclusion of three colour systems providing the primary subtractive colours of yellow, magenta and cyan which are mixed to produce all the other colours. This together with the higher cost of the internal processing systems makes them 2–5 times more expensive than pen plotters of equivalent size.

Pen plotters are vector devices that draw lines between coordinates in any direction by using x and y axes movements simultaneously. The axes movements are obtained by moving either pen or paper, depending on the model. Lines and text are drawn with fibre-tip, wet-ink or ball-point pens which are automatically lifted where no line or text is needed. Pen acceleration and maximum speed should be considered when selecting a plotter as these will dictate the throughput rate and also influence line quality. Standard quality plotters have a resolution of 0.025 mm, whilst expensive models can give 0.00015 mm. A variety of pen colours and thicknesses is available. Both cut sheets and rolls of drawing material are used. A4 and A3 plotters are usually the desk-mounted flat-bed type in which the bed may be horizontal, vertical or inclined. Larger sizes are usually floor mounted and may be flat-bed or drum type.

Flat bed pen plotters hold the paper fixed to a flat surface. A number of pens (usually 4 to 12) is held in a magazine and selected by the head when it is programmed to do so. Pen movement is obtained by either moving the pen head in both x and y directions or, alternatively, by moving the pen in only one direction with movement of the bed providing the other axis.

Drum pen plotters have the paper fixed to a drum which is horizontal. The drum rotates back and forth to provide one axis of movement, allowing the paper that is not in contact with the drum to hang vertically. Pens move along the axis of the drum to provide the other direction of travel. Because one axis relies on the rotational accuracy of the drum, they are not as accurate as flat-bed plotters. Drum plotters have the advantage of taking less floor space than flat-bed plotters because the drawing medium hangs vertically.

Electrostatic plotters are raster devices that produce an electrostatic charge onto a dielectrically coated medium via rows of styluses set in a write head. A raster scan technique is used to deposit small electrostatic dots on the medium. This then passes through a toner section in which particles of the toner powder attach to the charged medium. The image is then fixed by heat. The medium is drawn off reels of various widths and fed through the plotter. Plotting is very fast with possible throughput rates 5–20 times as fast as that of pen plotters. However, they typically cost 2–4 times as much as pen plotters of comparable size.

Thermal plotters use a process of applying wax-based ink onto the print medium by a combination of heat and pressure. They use a full page raster technique and are typically faster than a pen plotter. Resolutions of 300 and 400 dpi are achieved. The majority of machines are monotone, colour being substantially more expensive. Thermal plotters have few moving parts and, consequently, are quieter than pen plotters.

Printers In *laser printers/plotters* a laser is used to project a light beam onto a mirror which deflects the beam onto the paper. Rotation of the mirror provides movement of the beam across the paper. The paper is supported by a drum which rotates to provide movement from one line to the next. The drum is photosensitive and the image is built up pixel by pixel. Laser printers have gained in popularity because, with resolutions of 200–400 dpi, the quality level of the images is superior to that of dot-matrix-printer output. They are faster than dot-matrix printers, particularly for multiple copies, and much quieter, but they are more expensive.

Dot matrix printers are one of the most widely used devices for producing hard copy from computer systems and are used for text and low-resolution line work. An array of mechanical pins is used, each pin hitting an ink ribbon to deposit a dot of ink on the paper. The most usual pin-array pattern is 9 vertical by 5 horizontal giving a rectangle of 45 pins from which to produce the desired character. Selection of pin usage by the software can create a wide range of characters and symbols in a variety of font styles. The printer prints a line from left to right and at the end of the line automatically returns the carriage and advances the paper. Better images can be created by overprinting on the carriage return right to left or by the inclusion of more pins in the print-head design. A second set of pins can be offset slightly to produce a greater density of print on the paper.

14.5 Numerical engineering and control

14.5.1 Introduction

14.5.1.1 Definition of terms

There are a number of acronyms and terms used in this field which are commonly understood to mean:

Numerical engineering (NE)—a general term used to cover all numerically based engineering activity.
Numerical control (NC)—control of a machine by the repetitive reading of a paper or magnetic tape.
Computer numerical control (CNC)—control of machine by its own resident computer.
Direct numerical control (DNC)—the control of a number of machines by a remote computer which communicates with each machine's resident computer.

14.5.1.2 The advantages of numerical engineering

The development of numerical engineering and control technology has been a major factor in the improvement of manufacturing efficiency over the last 30 years. NC, CNC and DNC machines have now largely replaced manually and electromechanically controlled equipment and have been applied to many machining, forming, assembly and inspection processes.

NE has greatly improved productivity, quality and flexibility in manufacturing processes by providing machines that have the following advantages.

15

Manufacturing and Operations Management

J L Burbidge

Contents

15.1 Introduction to manufacturing management 15/3
15.1.1 Management processes and functions 15/3

15.2 Types of production 15/4
15.2.1 Material conversion classification 15/4
15.2.2 Classification of type of market 15/5
15.2.3 Combined classification 15/5

15.3 Systems theory 15/5
15.3.1 Types of variable 15/6
15.3.2 Feedback control 15/6
15.3.3 The design of the system 15/7
15.3.4 Connectance 15/7

15.4 Management and the functions 15/7
15.4.1 Management of functions 15/7
15.4.2 Hierarchical levels 15/8
15.4.3 Integrated and computer-integrated manufacturing 15/11
15.4.4 Recording a management system 15/12

15.5 General management 15/12
15.5.1 Strategy and policy 15/12
15.5.2 Organisation 15/13
15.5.3 Personnel and general control 15/13

15.6 Product design 15/14
15.6.1 The product specification 15/14
15.6.2 Design modifications 15/14
15.6.3 Quality control 15/14

15.7 Production planning 15/15
15.7.1 The planning specification 15/15
15.7.2 Make or buy 15/15
15.7.3 The choice of material form 15/16
15.7.4 The choice of processing route 15/16
15.7.5 Planning tooling requirements 15/17
15.7.6 Work study 15/18

15.8 Design of the material flow system 15/18
15.8.1 Organisation at workshop level 15/18
15.8.2 Continuous line flow 15/19
15.8.3 Group technology—component processing 15/20
15.8.4 Classification and coding 15/22
15.8.5 Production flow analysis 15/23
15.8.6 Factory flow analysis 15/23
15.8.7 Group analysis 15/27
15.8.8 Group analysis—the division into modules 15/27
15.8.9 Finding groups and families 15/29

15.9 Assembly groups 15/32
15.9.1 Planning assembly groups 15/32
15.9.2 Traditional assembly organisation 15/32

15.10 Marketing 15/35
15.10.1 The long-term sales programme 15/35
15.10.2 The series of short-term programmes 15/38

15.11 Production control 15/38
15.11.1 Programming—general 15/40
15.11.2 The annual programmes 15/42
15.11.3 Flexible programming 15/42
15.11.4 Programme reconciliation 15/43
15.11.5 Ordering 15/43
15.11.6 Single- and multi-cycle ordering 15/43
15.11.7 Just-in-time production control systems 15/46
15.11.8 Stock-based ordering methods 15/46
15.11.9 Flow control ordering methods 15/48
15.11.10 Dispatching 15/50

15.12 Purchasing 15/52
15.12.1 Batch buying 15/52
15.12.2 Schedule buying 15/53
15.12.3 The call-off method 15/53
15.12.4 The choice of supplier 15/53

15.13 Finance 15/54
15.13.1 Cost accounting 15/54
15.13.2 Break-even analysis 15/54
15.13.3 Conclusion 15/55

15.14 Personnel 15/55

15.15 Secretarial function 15/55

15.16 Conclusion 15/56

15.1 Introduction to manufacturing management

In the early 1900s, the dictionary defined 'manufacturing' as 'the making of wares by hand or machine', and defined 'production' as 'the manufacture and distribution of goods'. Manufacturing was then, and even with the changes in its meaning with time, is still, that part of production which is concerned with making the product.

It is impossible, however, to make an intelligent study of the making of products without some consideration of the markets for which they are made. The distinction between making products (in the factory) and distributing them (in the market) is not therefore a useful one, for those who wish to study either production or manufacturing in depth.

Today, without any attempt to redefine the difference between the terms 'production' and 'manufacturing', the latter has become the fashionable word. It is proposed here to treat the two words as synonyms. The distinction between them was made by the early economists; it may still have some value in the science of economics, but it has little or nothing to offer to the science which we are studying. Manufacturing management is, therefore, a large part, or perhaps all, of the science of 'production management'. It is an applied science which draws on most of the other sciences. In this chapter we will look mainly at those functions concerned with the making of products and will deal only briefly with the remaining functions.

15.1.1 Management processes and functions

Management can be defined as: the art and science of planning, directing and controlling human effort, so that the objectives of an enterprise may be attained. It can be seen as a set of thousands of different tasks devoted to this objective. The processes of management are listed below.

(1) *Planning*, or deciding what to do in the future.
(2) *Direction*, or 'the process of management by which plans are caused to be implemented'. Direction covers both the issue of orders and the means used to persuade people to carry them out.
(3) *Control*, or 'the process of management which constrains events to follow plans'. Control includes monitoring which compares plans with actual output, and feedback which transmits information about significant variances between plans and output to management, so that action can be taken to correct them.

All management tasks require the application of all three of these processes for their successful completion.

Management tasks can also be sorted into sets of closely related tasks, which require similar skills for their efficient performance. These sets of tasks are known as 'management functions' and form the basic classification of the science of production management.

Table 15.1 lists the eight main management functions, describes the types of tasks which they include, notes the types of skill needed for each of them, and lists some of their principal inputs and outputs, and the special controls used with each of them.

Although the functions are partly autonomous, including many tasks which are of no interest outside their own function, the functions are nevertheless closely related: first because, for the efficiency of the enterprise as a whole, they need to share common goals and objectives; and second, because a part of the data input needed by each function comes from

Table 15.1 The management functions

*Function**	*Type of task*	*Controls*	*Inputs*	*Outputs*
Product design (design)	Plans final form of product	Quality control	Ideas; market research; R&D	Parts list; drawings
Production planning (production engineer)	Plans how product is to be made	Process control; maintenance	Parts list; drawings; sales forecast	Plant list; layout; operating times; tooling and routes
Production control (work scheduling)	Plans material supply and processing activities	Progressing; loading; inventory control	Sales progress; parts list; plant list; routes; operating times	Programmes; orders; purchase delivery schedules
Purchasing (commercial)	Finds sources; makes supply contracts	Purchase progressing	Purchase delivery schedules	Purchase orders
Marketing (commercial)	Finds/develops markets; sales and distribution	Sales control	Sales orders	Sales records; sales programmes
Finance (financial)	Plans investment, profit and cash flow	Budgetary control; standard costing	Annual programmes; invoices *in/out*; bank payments in/out	Budgets; accounts; balance sheet; profit and loss account
Personnel (people)	Plans employment; conditions; welfare; training; promotion	Merit rating; attendance	Hired; fired; promoted; retired	Employee list; conditions of employment
Secretarial (data processing)	Plans communications data process and store	Data control	Production system design	Software

* The main skills required in each function are given in parentheses.

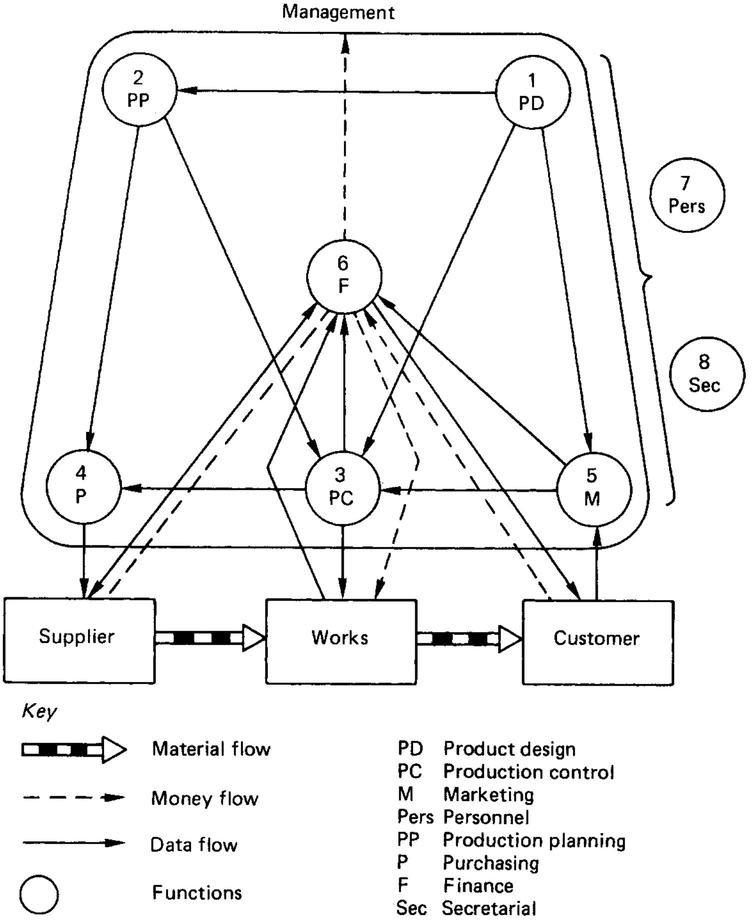


Figure 15.1 The arterial flow system

some other function or functions, and a part of their output will again provide the input for some other function or functions, or provide instructions to the works, the company suppliers or its customers.

Figure 15.1 shows the main information flow between the functions, by means of a network diagram. This primary flow of information is called the 'arterial flow system'. Note that this chart also shows the flow of information between the management functions, the works, and the company's customers and suppliers, and shows the principal money flow routes.

15.2 Types of production

There are many different ways in which production enterprises can be classified by type. They include, for example:

(1) classification by type of product produced, e.g. boats, machine tools, aero engines;
(2) classification by volume of output, e.g. mass production and job lot; and
(3) classification by the amount of capital employed, e.g. capital intensive, and manual or craft.

For manufacturing management, however, we need a classification that brings together factories which are faced with the same types of management task. For this purpose, a two-part classification has been adopted which combines a material conversion type classification, based on the ratio of material varieties to product varieties, and a classification by the type of market served (i.e. large contract, jobbing, batch or continuous flow).

15.2.1 Material conversion classification

Figure 15.2 shows the material conversion classification of industries into four main types and gives examples of each of them. Process industries are industries in which a small number of material varieties are converted into an equally small number of product varieties. Material conversion is generally concerned with a small number of processes always used in the same sequence.

Implosive industries manufacture parts and are those in which a small number of material varieties are converted into a large number of part varieties. Like process industries, they usually use a small number of processes in the same sequence, but one or two of these processes divides the material flow into many different components. The foundry industry is a typical

	Process	*Implosive*	*Square*	*Explosive*
Material input Product output Material flow type	Bulk material Bulk material Line flow	Bulk, or general material Components Batch/line flow	Components Components Batch flow	General or special materials Assemblies Batch flow
Examples	B Cement E Ore treat (mine) F Milk F Sugar F Distilleries C Gases (oxygen, nitrogen) B Bricks B Timber F Breweries Tanneries Paper	E Foundries D Potteries D Glass C Decorative laminates T Spinning, fibres E Brake linings Printing T Knitting F Bakeries E Rolling mills E Wire drawing	E Jobbing machine T Dyeing textiles T Finishing textiles E Heat treatment X-ray E Painting E Electroplating E Metal spraying Polishing	E Automobile E Electronics D Consumer durables E Machine tools E Internal combustion engines E Electrical T Weaving D Clothing and shoes C Chem. dyestuffs D Furniture E Welding

Figure 15.2 Types of industry according to the material-conversion classification. *m*, Number of material varieties; *p*, number of product varieties. Industry types: F, food; C, chemicals; T, textiles; E, engineering; B, building; D, domestic

example: the main materials are sand, coke and pig iron; the processes are metal melting, moulding, pouring, knocking-out and fettling; diversification to produce a large number of different castings takes place at the moulding stage.

Square industries also manufacture components and are those in which a large number of material varieties are converted into an equally large number of component varieties. Square industries are often service industries which provide a special service to other types of industry.

Explosive industries are those which convert a large number of material varieties into a relatively small number of product varieties. These are assembly industries which combine components to produce assembled products.

15.2.2 Classification by type of market

The second type of classification used in this chapter is based on the type of market served. Again there are four main classes:

(1) large contract (non-repetitive),
(2) jobbing (non-repetitive),
(3) batch (repetitive), and
(4) continuous (repetitive).

The large-contract type covers non-repetitive products such as ships, special machines, power stations and new factories. The regulation and control of the production of such products is a type of project management. It uses such techniques as critical path analysis (CPA), or the programme evaluation and review technique (PERT).

The term 'jobbing' is generally reserved for non-repetitive orders for small parts or assembled products which are made in factory workshops. Because the processes are non-repetitive they are usually 'made-to-order' products. It is possible that some of these items may be ordered again later, but as there is no certainty that this will happen, they can only be treated as non-repetitive.

Batch-production products are repetitive products which are made intermittently in batches. Such products generally share the capacity (machine hours) of the machine tools or other production facilities used to make them with other products in the product range.

Continuous-production products are repetitive products which are made on a continuous flow line. The machines and other processing facilities are laid out in a line in the sequence in which they are used, and are reserved for the manufacture of one particular part, or of a 'family' of similar parts.

15.2.3 Combined classification

The above two classifications, by material conversion type and by type of market, can be combined to give 16 classes, as shown in *Figure 15.3*. Four of these classes (CP, CI, CS and JP) are trivial, being very unlikely to occur in practice. Three others (LI, LS and LE) are unlikely to be found in the pure form, but may be found in partial form combined with some other class. The LE class, which in the form of the automated transfer line, for some years appeared to be the ultimate objective for explosive products, has lost ground to the BE (batch/explosive) class, and its automated form of the flexible manufacturing system (FMS), due to the growing demand of markets for a wider choice of product varieties.

Figure 15.3 also shows the form of material and the form of product produced by each type of industry. In particular, only the explosive industries produce assembled products, whilst the remainder produce either bulk products (measured in general units of weight, volume, area or length) or components.

15.3 Systems theory

Systems theory provides a means for studying complex systems. It has been particularly useful for the study of manufacturing systems.

A system can be defined as: a set of variables which together form an easily identified and definable entity, and are so related that a change in any one of them will induce, or be induced by, a change in at least one of the others. A manufacturing system is then a set of all those variables involved in production, which are so related that changes in their values cause the values of some of the other variables in the set to change.

	(P) = Process m ═══ p	(I) = Impulsive m ◁ p	(S) = Square m □ p	(E) = Explosive m ▷ p
(C) Contract	CP	CI	CS	CE m = components p = a complex assembled product
(J) Jobbing	JP	JI m = bulk materials p = components	JS m = components p = components	JE m = components p = assembled products
(B) Batch	BP m = bulk materials p = bulk materials	BI m = bulk materials p = components	BS m = components p = components	BE m = components p = assembled products
(L) Line	LP m = bulk materials p = bulk materials	LI m = bulk materials p = components	LS m = components p = components	LE m = components p = assembled products

Figure 15.3 Combined classification of production. *m*, Number of material varieties; *p*, number of product varieties

15.3.1 Types of variable

The variables in any system can be divided into four main types: parameters, output, input and system design.

Parameters Also known as *regulatory parameters*, these are variables to which a manager can assign arbitrary values at will. Examples from manufacturing are order quantity, run quantity and selling price. Within certain limits, a manager can assign any value to such variables, and this will be their real value. The manager may be influenced in his choice of parameter values by his knowledge of the value of other variables, but changes in the values of these other variables have no direct effect on the value of the parameters.

Output variables These are variables to which a manager cannot assign arbitrary values. He can only change their values indirectly by changing the value of appropriate parameters. Examples of output variables are cost, stock, profit and scrap. In the case of stock, for example, the manager cannot change the stock value directly. He can, however, reduce the stock volume by increasing the value of the run frequency parameter (runs per year) which at constant output rate will also reduce the value of the 'run quantity', or number of parts made together as a batch before changing to make some other part.

Input variables Also known as *uncontrollable input variables*, these are variables imposed by the environment in which the enterprise operates. The manager has little chance to influence the values of these variables, but they do have an important effect on the values of the output variables. Examples are the weather, the exchange value of the pound (£), and the rate of interest charged on loans.

System design parameters These are parameters in that a manager can assign arbitrary values to them at will. In this case, however, the choice for each variable is between two alternative values, unlike the regulatory parameters where the manager chooses a value from a range of values. Examples are make or buy, multi- or single-cycle ordering, and process or product organisation. The design of a system can be seen as the choice between large numbers of alternatives such as these.

15.3.2 Feedback control

Figure 15.4 is a diagrammatic representation of a system. The rectangle represents the system, and the arrows leading into and out of it represent the different types of variable which influence its performance.

The same figure also shows the feedback control loop, with which the manager tries to control the output variable values. It is assumed that the manager has planned a series of periodic values for the output variables which he wants to achieve. He

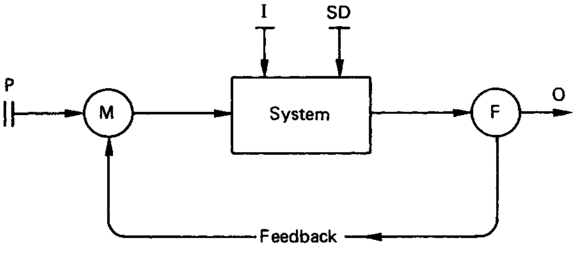


Figure 15.4 Diagrammatic representation of a system. P, parameter; O, output variable; I, input variable; M, manager; F, monitoring; SD, system design variable

arranges for the actual values of these output variables to be measured at the end of each period, and for these values to be compared with the planned values. This process is known as 'monitoring'. Details of any significant variances between the plans and actual output values are fed back to the appropriate manager, who either changes parameter values to bring output values back into line with the plans, or changes the plans for the same reason.

In simple terms, the manager regulates the output of the system by applying the management processes of planning, direction and control. For example, in the case of planning the output of finished products from the factory, the manager first 'plans' a 'production programme' showing the number of products to be made in a series of future time periods. He then 'directs' that these plans should be implemented by issuing the 'production programme' as an order to the factory. Finally, he 'controls' to ensure that the plans are followed by 'monitoring' and by arranging that he is informed (feedback) if there are significant 'variances' between the planned and the actual output value.

The difficulties with this simple concept are: there are hundreds of output variables; the output variables interact in that the values achieved for one variable will affect the values of some of the others; and these relationships are extremely complex and difficult to forecast. It can be said that the production system is very complex, and that it comes into the class of cybernetic systems, or very complex systems, which we can never hope to understand in all their infinite detail.

15.3.3 The design of the system

Even though we may not understand the minute detail of how a production system works, we can learn to predict how it will behave in given circumstances, and how to steer the output variable values in a required direction.

The illustration in *Figure 15.4* was first drawn by a man called Mr Black and the rectangle in this figure is, therefore, widely known as the 'Black box'. Black showed that if we measure at regular intervals the values of the parameters, uncontrollable input variables and output variables, we can learn to anticipate how the system will react to different parameter changes, even though we do not understand in detail the way in which the system works.

We do not, however, have to accept the production system as we find it. We know that we can improve its performance by improving its design. We know, for example, from experience and from experiment, that some types of system are more 'sensitive' than others. In other words, a change in the parameter values in some systems induces a change in related output variable values much more quickly than in others. We also know that certain types of production system are more 'predictable' than others. In other words, there is a high probability that a given change in parameter values will induce a predictable change in the values of some output variables.

Thus we know that we can learn to design production systems which are very much more efficient than those being used at present. As was mentioned earlier in this section, we can see that system design requires a choice between a very large number of alternatives.

15.3.4 Connectance

The main outstanding problem in systems theory is: How can one learn to understand the relationship between the different variables in the system?

The author has postulated that it is impossible to make general statements about the quantitative relationship between production system variables, because these depend on the nature of the enterprise, on the design of its system, and on the values of the uncontrollable input variables. We can, however, make general statements about the direction of the change in value of one variable which is induced by a given direction of change in another related variable. We can hope, therefore, to learn how to steer system output variable values in required directions, even though we may be unable to predict the quantitative effects of such changes.

The author has catalogued several hundreds of these direction-of-change relationships in order to form a relational data bank. *Figure 15.5* for example, shows the variable 'run quantity', i.e. the number of parts of the same design which are run off at a particular work centre before changing to make some other part. The figure shows that reducing the run quantity at constant output rate, reduces

(1) cycle stocks,
(2) thoughput time, and
(3) stock holding costs.

It will be realised that changes in these output variables themselves induce further changes in other variables, which in turn induce even more changes. *Figure 15.6* shows the results achieved using a computer program called 'Cascade', which explores these sequential types of change.

Figure 15.7 illustrates the results achieved with a computer program called 'Cure', which reverses the Cascade solution, and says: I have this problem; what changes must I make to solve it?

15.4 Management and the functions

Manufacturing is concerned with the conversion of materials into products for sale to customers. This simple idea is illustrated in the diagram in *Figure 15.8*.

Production management is, directly or indirectly, concerned with the regulation and control of the flow of materials into the works from suppliers; between conversion processes inside the works, and finally of finished products to the company's customers. The hundreds of different management tasks involved in this work in any factory can be classified according to:

(1) function, and
(2) hierarchical level.

15.4.1 Management of functions

The division into management functions is listed in *Table 15.1*. These eight functions provide the basic classification of the science of production management.

For the study of management tasks and of the way in which they are related, the division into functions provides an ideal starting point. The majority of management tasks are specific to only one function. This makes it possible to divide management into largely independent specialised 'functions', or 'subjects for study'.

Difficulties with the functional approach only arise when we try to organise the work of management, or in other words to allocate the responsibility for different sets of tasks to people in factories. Traditionally, in most factories the management was organised functionally, with different departments for e.g. product design, production planning, production control, purchasing, marketing, finance, personnel and secretarial work. Although the work of these functional departments was loosely coordinated by the general management, in practice they operated largely as independent kingdoms.

	Regulatory parameter	P	3	0	3
1.	Name: Run qty – Mfgr. (R.Q)				
2.	Definition: The quantity of a part run off at a work centre before changing to produce some other part				

3. Function:

1.	Product design		6.	Finance			↑					
2.	Production planning		7.	Personnel			|					
3.	Production control	✓	8.	Secretarial			3					
4.	Purchasing		9.	General Manager								
5.	Marketing		0.									

4 Limitation:

											L
1.	Run frequency	↑ ↓	fv	P	3	0	4	3	2		
2.	Transfer quantity	↑ ↑	max.	P	3	0	5	5	1		
3.											

5. Induction:

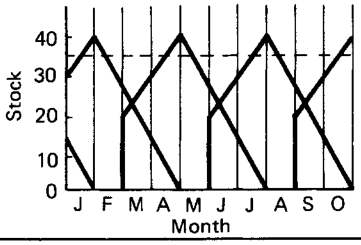
									C
1.	Stocks	↑ ↑	0	3	4	0	5	1	0
2.	Cycle throughput time (Mfgr)	↓ ↓	0	3	3	0	1	1	0
3.	Cost – stock holding	↑ ↑	0	6	8	2	0	1	0
4.									
5.									

6. Notes

Stock charts demonstrate induced changes

(1) 4 runs pa of 42 = 168
Average stock = 36

(2) 12 runs pa of 14 = 168
Average stock = 12



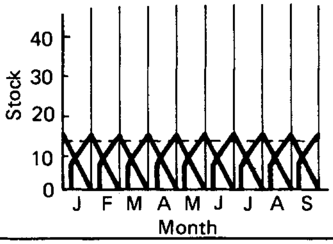


Figure 15.5 A regulatory parameter. Mfgr, manufacturing; fv, a direct function of the variable under consideration; pa, per annum

This type of organisation is strong in the coordination of tasks specific to particular functions, but weak in coordinating tasks which are common to several functions, and in coordinating functional plans to achieve the best overall company results.

15.4.2 Hierarchical levels

There is a general need in management for some form of hierarchical control. In other words, there is a need for levels of management, starting at the top with direction or general management which is responsible for strategy and the regulation and control of overall performance. One or more further levels are needed to elaborate the plans of general management in greater detail, leading to the supervisory level which is responsible for the regulation and control of the direct workers.

Figure 15.9 shows a company with three levels of management—general management, functional or middle management, and supervisory management—in the form of a pyra-

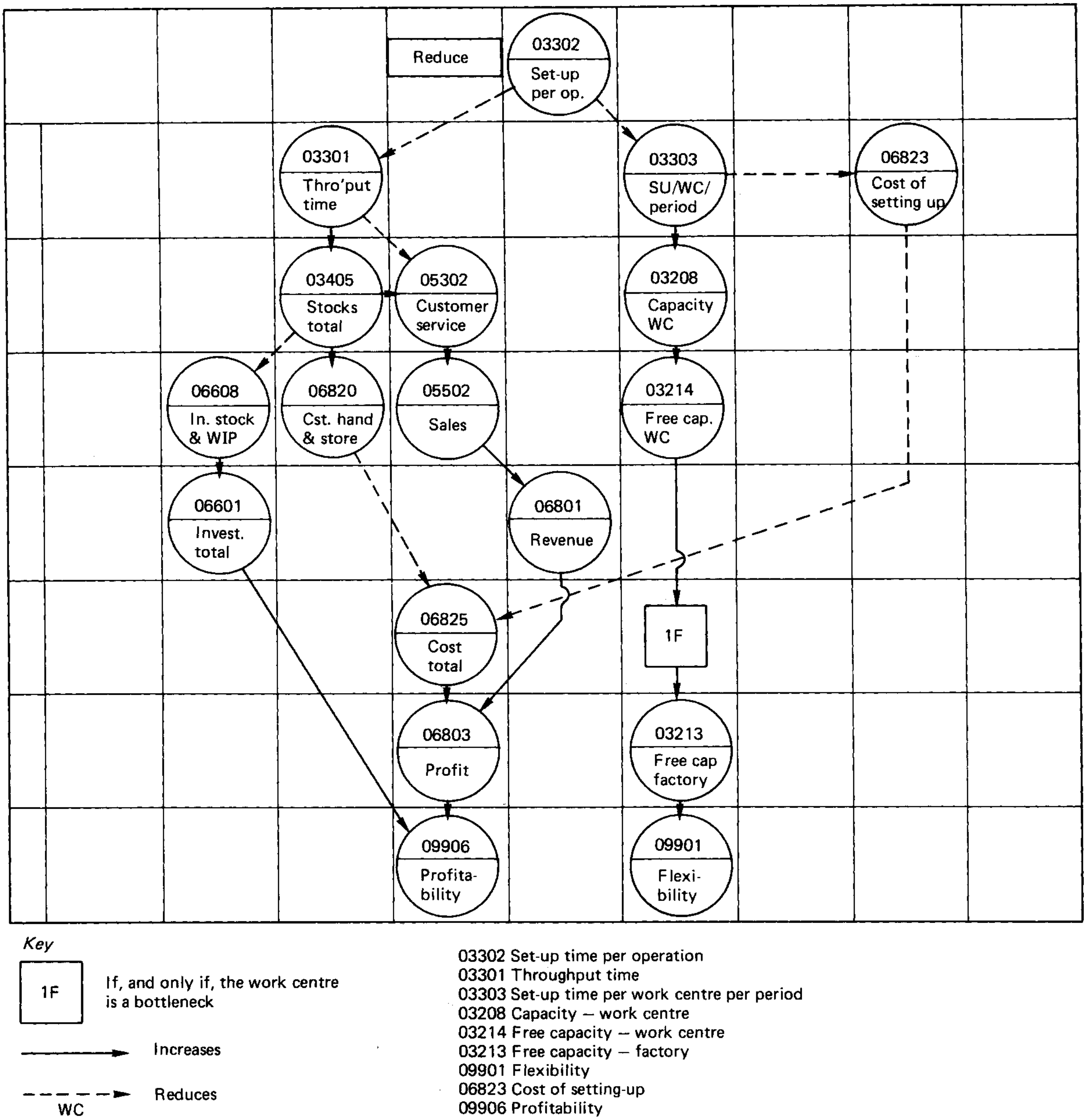


Figure 15.6 A cascade network showing the effect of reducing the set-up time per operation

mid. The base of the pyramid is the material flow system of the company. It shows machine tools and other work centres and the way in which materials flow between them during production. At each of the levels in *Figure 15.9*, a number of decision-making tasks must be carried out. These are connected with other tasks, both at the same level and also at other levels. The output from each task provides either an output from the system or an input to some other task inside the system. This representation of management as a pyramid is borrowed from the GRAI method, developed by Professor Pun and Dr Guy Doumeingts of Bordeaux University. They have developed methods for planning management systems, based on the hierarchical decision-making flow illustrated in *Figure 15.9*.

The GRAI method, and other less formal methods which plan management systems hierarchically, tend to weaken the functional divisions in management, and to strengthen the central control of the enterprise as a whole and the connection

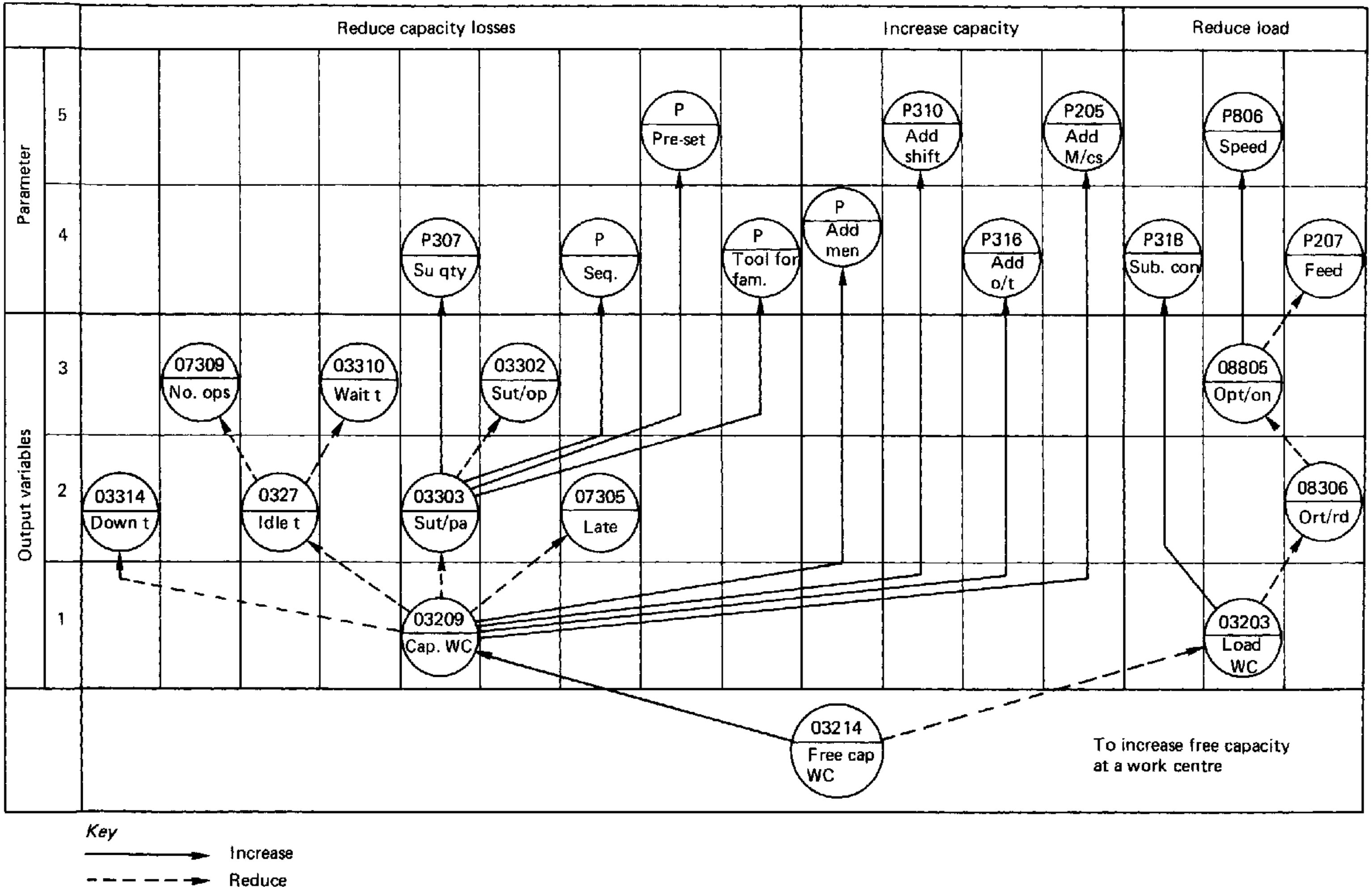


Figure 15.7 A cure network

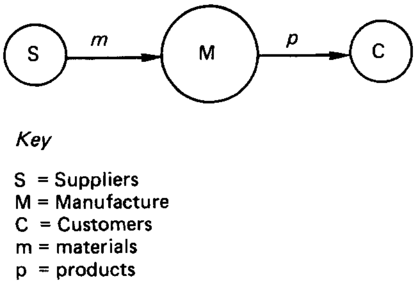


Figure 15.8 Manufacturing

between decisions made at different levels. Production control, for example, is the scheduling function in management. It operates at the levels of:

(1) *programming*—schedules the completion of finished products for sale;
(2) *ordering*—schedules the completion of made parts and the delivery of purchases; and
(3) *dispatching*—schedules the completion of operations in the manufacture of parts.

With the traditional functional organisation, all of this work would be done in a production-control department. With hierarchical planning, there is a tendency for top-level decisions to be made at the top or general management level, and for lower level decisions to be delegated to the lowest, or supervisory level of management.

15.4.3 Integrated and computer-integrated manufacturing

The functional organisation of management was simple to organise and to understand, but its failure to consider the enterprise as a single whole was a fundamental weakness, which cannot be tolerated in the future. Industry cannot be run efficiently by a collection of independent warring kingdoms.

The problems of the need for change in management organisation have been reduced by the development of computerised data processing. Once the strategies and policies for management have been fixed, and the primary operating plans have been agreed, what is left is systematic data processing, which can be automated using a computer.

It is still true that most management tasks are specific to one function only, even if we do not accept the functions as the basis for management organisation. The core of the computerised data-processing system is, therefore, a number of subsystems covering functional needs. The output of these subsystems forms the input to some other subsystem, or provides data for suppliers, the 'works', or customers. These major data flows form the arterial flow system (see *Figure 15.1*) which can also be automated.

Every manufacturing enterprise is unique, and thus the management system of each enterprise must also be unique. What industry really needs is tailor-made computer software, not standard operating packages. In the early days, computer operating system software was sold for universal or wide application. It was generally sold on the basis that the software program could be amended to suit the special needs of the customer. In practice, because these programs were written in first-generation languages such as Fortran, they were very

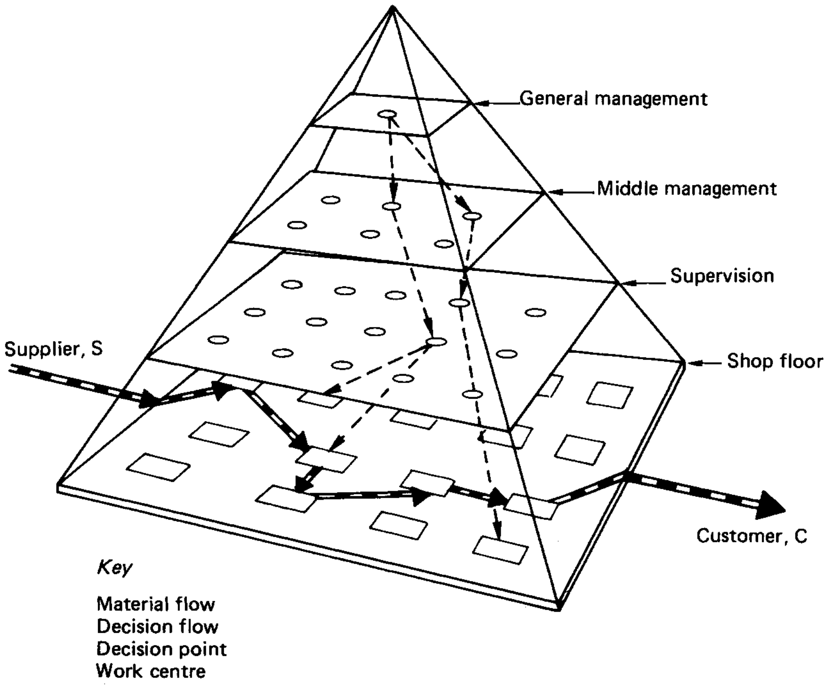


Figure 15.9 Hierarchical levels of management

difficult to change. Companies ended up using only a small part of a very expensive package, or trying to change their management system to suit the package. Present-day packages are not much better, but there is some hope for the future in the latest 'distributed databases' which, if used with fourth-generation computer languages, make it much simpler for companies to write their own programs.

The first need is to design an integrated manufacturing (IM) system in which: each task is done only once (data-processing tasks are not repeated in several functional subsystems); data are stored in one place only (but not necessarily all together in the same place); and the subsystems have been simplified as far as possible by integration to eliminate unnecessary variety.

It is only when the IM system has been designed that computerisation to produce a computer-integrated manufacturing system (CIM) can be profitably considered.

15.4.4 Recording a management system

Because management systems are generally very complex, a major problem is how to record such systems so that they can be checked efficiently, and can be communicated to those (e.g. software designers) who need to understand how they work, without risk of ambiguity.

One widely known method for recording system design is the IDEF system developed by the USA Space Agency. This is too complex for general use in management, and there have been several attempts to simplify it for this purpose. One of these, known as MODUS (developed by BIS Applied Systems) is illustrated in *Figure 15.10*. This figure shows the first level division of a production-control system into the progressive subsystems of programming, ordering and dispatching, together with the three main feedback control systems of progressing, loading and inventory control.

As an expansion of *Figure 15.10*, *Figure 15.11* shows the subsystem of 'ordering' in detail, to show how it works.

The MODUS method uses progressive charts such as these to illustrate a management system. The method of drawing the charts is obvious, but it might be mentioned that they are easier to read if the main data flows are from left to right, as shown in *Figures 15.10* and *15.11*, and only 'feedback' is in the reverse direction.

15.5 General management

General management is the management of the enterprise as a whole, and is the work done by the general manager or chief executive of the enterprise. He exercises direct authority over the senior managers—those of marketing, design, manufacturing, finance, personnel, etc.—and is directly responsible for strategy, the main lines of policy and policy coordination, organisation, personnel, and the general control of operations.

15.5.1 Strategy and policy

'Strategy' is a word borrowed from military science, which means 'generalship'. It covers the first level of planning in an enterprise, including the fixing of objectives, the organisation or the allocation of responsibility for different types of task to individuals, and the specification of general lines of action to be followed in order to achieve the selected objectives.

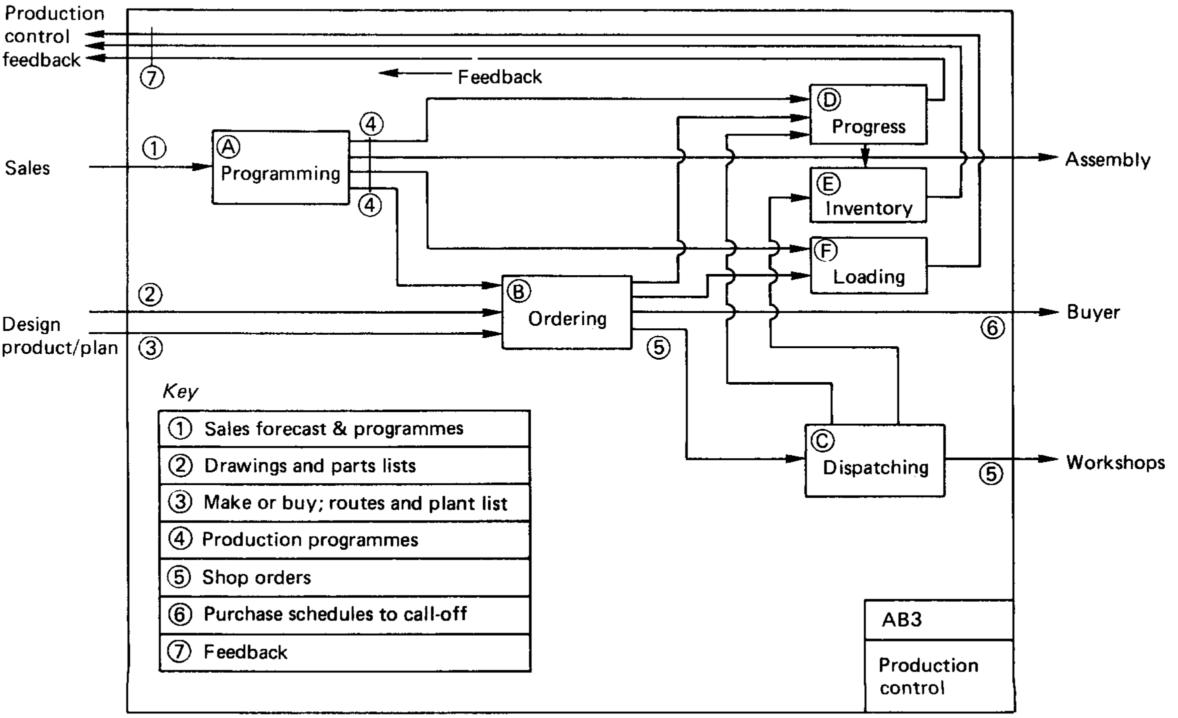


Figure 15.10 Production control

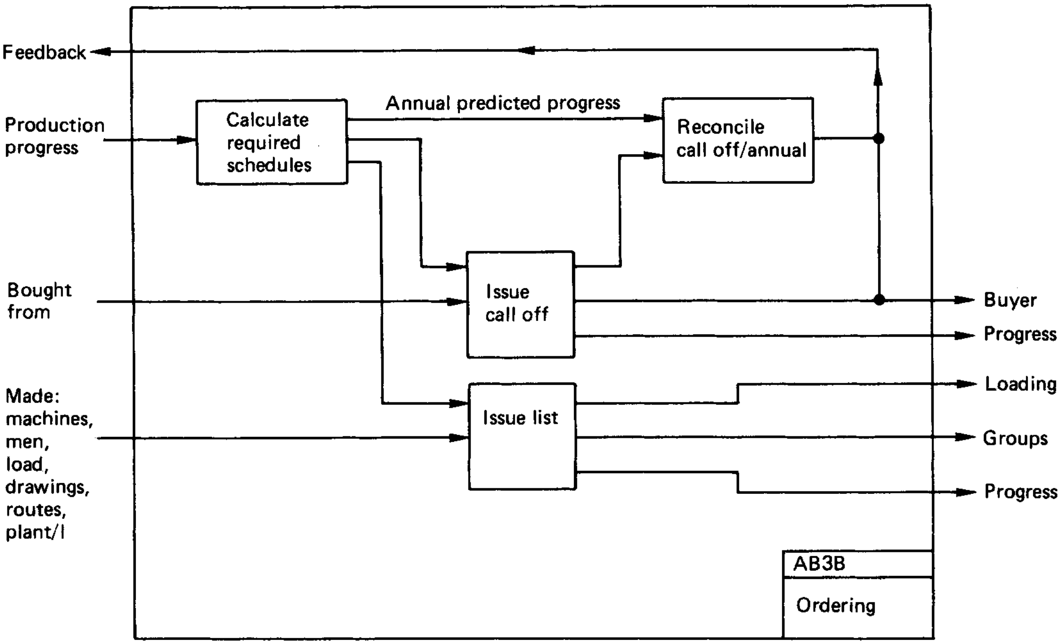


Figure 15.11 Ordering with period batch control (see Section 15.11.9.2). N.B. with period batch control, each product in the programme requires: (1) a standard number of parts to be made in each group; (2) a standard load in machine hours on each machine in each group; and (3) a standard number of parts to call off from each supplier

'Policy' can be defined as a plan of action. In military science, strategy is seen as management directed towards the winning of a war, while policies are concerned at a lower level with the winning of battles. For example, in manufacturing a strategy might be adopted: 'to minimise the stock investment' and policies which could be adopted to this end might include the introduction of group technology and just-in-time (JIT) production control.

A general manager may delegate the responsibility for the lower levels of policy, but he must coordinate its introduction to ensure that his managers do not introduce policies which conflict with the strategies he has adopted, or with each other. The strategies will generally be fixed at an overall company level; the policies are generally concerned with different functions.

15.5.2 Organisation

Organisation can be seen as the bringing together of tasks into sets, and the allocation of responsibility for those sets of tasks to people. Such allocations must generally be hierarchical. In other words, the general manager cannot directly control all the people in an enterprise. He will be responsible for the senior managers in the enterprise, and they in turn will have others reporting to them at the next level.

The type of organisation adopted by an enterprise will depend on the strategies and objectives which it adopts. Management organisation in recent years has been strongly functional, with largely independent departments concerned with each of the eight functions. This type of organisation is now changing. Strategies which emphasise the importance of flexibility—or the ability rapidly to follow changes in market demand, without losses due to materials obsolescence—have tended to change organisational structures towards cross-functional, multidisciplinary forms, giving better coordination.

Again, in the past most workshops in manufacturing enterprises used process organisation. New strategies such as low stock manufacturing, total quality control, increased worker participation in decision-making, and flexibility, have forced the change from process to product organisation, and in particular towards group technology where organisational units complete 'families' of products or parts.

15.5.3 Personnel and general control

It has been said that the most important asset of any enterprise is its people. It is the major responsibility of any general manager to demonstrate his belief in this idea. Successful management styles differ greatly from one general manager to another. There is no standard approach which should be adopted by all managers. The main need is to show that he is interested in the people in the enterprise, and that he notices what they are doing and has some understanding of how they feel.

The main contact of the general manager of an enterprise will be with the senior first-line managers who report to him. To maintain control of the enterprise, he needs to monitor their performance, but he must try to do so in a way which will not kill their initiative and motivation.

To maintain full control and to coordinate decision-making in the different functions, the general manager will normally need to have direct control of the first level of decision-making in each function. In the product design function, for example, he should have the final word in the composition of the product range. In finance, the annual budget should be issued in his name, and in production control he should approve the

annual programmes before they are adopted. If he is to build an efficient and effective team, however, he needs to delegate the further levels of decision-making to his subordinates.

15.6 Product design

The product design function includes all management tasks which are concerned with planning, directing and controlling the final form of the products produced by a manufacturing enterprise. This work requires an element of creativity, coupled with experimental verification and improvement through research and development (R&D), and with quality control (QC), based on comparison between the actual results achieved in manufacture and the required measures shown in product specifications.

Product design is simple to control in most process industries. It is closely associated with production planning and production control in jobbing and contract industries (see *Figure 15.3*), where product design is often a part of any sales contract. In the remaining types of industry which normally make standard products, the main problems in the product design function are: how to specify the product design; how to regulate modifications to the design, and how to 'control' the quality of the finished product. This section of this chapter deals only with these three problems.

15.6.1 The product specification

A product specification is a record of the decisions made when designing a product. According to the type of industry, it may take the form of a model, a pattern, a sample, a chemical formula, a series of tests or, in its most complex form as the specification for an assembled mechanical product, it may comprise dimensioned drawings, a parts list ('bill of materials' in the USA), and other supporting specifications.

Product specifications are needed in manufacturing as:

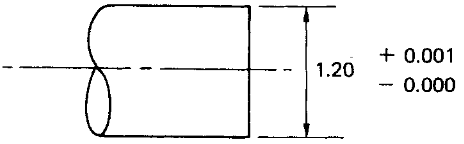
(1) the basis for sales contracts;
(2) the basis for purchase contracts;
(3) to provide the information needed by production planning to plan the methods of productions;
(4) the basis for quality control;
(5) the basis for production control; and
(6) the basis for computerised data processing.

For all these purposes, an accurate, unambiguous record which shows the required final form of the materials in the product is absolutely essential.

Consider the case of the specification for assembled mechanical engineering products. First, the drawings. Drawings are needed of the assembly as a whole, and also of each separate component included in the assembly. At least three views (or projections) are needed for each drawing—from the side, from the side at 90° to the left, and a view from overhead—to make them unambiguous.

The principal dimensions must be shown on the drawings in the form of maximum and minimum, or limiting, values. The difference between these 'limits' is known as the 'tolerance'. Limits are necessary because absolute perfection is impossible to obtain. The limits are fixed, (a) in order to achieve the required level of performance from the machine (operational dimensions), and (b) to ensure the interchangeability of components (see *Figure 15.12*).

The second requirement is a parts list. This must list every component included in the product. Each of these separate items must have its own name and also a code number which is



Limiting values: 1.201 and 1.2000
Tolerance: 0.001

Figure 15.12 Limits on drawings

specific to that item. Most engineering products can be divided into assemblies, subassemblies and so on. For convenience, in 'kitting out' or picking sets of parts for assembly, it is desirable that assemblies and subassemblies are shown separately in parts lists. This has the disadvantage that some items may appear more than once in the same parts list, but it is not difficult to produce consolidated lists of these items for use in component ordering.

The third specification requirement mentioned above, of other supporting specifications, may include material specifications, surface-finish specifications, and test specifications, describing the tests to be used to ensure that the machine is functioning satisfactorily.

15.6.2 Design modifications

During the life cycle of most products, there are occasions when the design of the product must be changed. It is desirable that the introduction of such 'modifications' should be carefully controlled, and that one should be able to tell which customers received products to the old design and which had the new modified design.

Modifications have different degrees of urgency. For example, if the change is needed to improve the safety of a product it may have to be introduced immediately in all products before dispatch, and even retrospectively in products which have already been sold. If, on the other hand, the modification is designed to reduce costs, then it is possible to wait for the best economic moment before it is introduced.

The retrospective change of products already sold is normally the most expensive form of modification. In other cases, material obsolescence is the most expensive cost induced by modification. When a modification is introduced, any parts or materials to the old design may only have scrap value. Where possible, the introduction of modifications should be arranged to allow time to use up old stocks. With traditional types of production control, this may require a wait of many months. It is one of the advantages of just-in-time low-stock production control systems that modifications can be introduced quickly without losses due to obsolescence.

A record is needed which shows the modification state of all products sold, so that the correct spare parts can be delivered if they are needed in the future.

15.6.3 Quality control

The feedback control system which constrains events to follow the plans contained in product design specifications, is called 'quality control'. The monitoring component of quality control which measures the actual dimensions achieved during manufacture compares them with the values shown on the drawings, and feeds back information about parts which are unacceptable, is called 'inspection'.

Systematic inspection is relatively new in industry. Before its introduction, quality was seen as the responsibility of the craftsmen who did the work, and they were expected to inspect their own work. Today, while it is accepted that there are parts of inspection requiring special skills (metrology, for example), there is a tendency to return to the old order. More and more companies are delegating inspection to the floor of the shop where the work is done, and using small quality-control units to check that the product is being inspected, and to check the work of the inspectors to ensure that the best methods are being used for inspection.

In the sense of manufacturing reliability or compliance with the product specification—measured in units of, say, inspection rejects per million parts produced—there is little doubt that for manual operations quality depends on the skill and motivation of the people who do the work. Inspection provides policing that eliminates items which are outside limits, but it does not generate an improvement in quality.

With the development of automation, however, the emphasis in manufacturing reliability is changing from control of the product to control of the process. If the process is controlled, then the product quality is also controlled.

In the present climate of high competition, quality control is more important than ever, but its scope is much wider than it used to be. Today, quality is measured in terms of customer satisfaction. Therefore, quality control is concerned not only with manufacture to specification, but also with design, the methods of manufacture, marketing, after-sales service and any other factor which affects customer satisfaction.

15.7 Production planning

Production planning is the function of management which plans, directs and controls the physical means used to manufacture products in the enterprise. It can conveniently be studied in two parts: (a) the design of the detailed methods to be used to make components and to assemble them to form products; and (b) the methods used to organise the work and to plan the material flow system and plant layout for the enterprise. These two aspects are described in this section and Section 15.8, respectively.

Before production planning can commence, the planners need two types of information. First, they need copies of the product specifications for all the different products to be produced. Second, they need a planning specification which fixes the criteria to be used in making the choice between alternative planning possibilities.

In the most complex form of production planning in explosive industries (e.g. mechanical engineering assembly), seven main types of decision have to be made one after the other for each component:

(1) make or buy,
(2) the choice of material form,
(3) the processing route,
(4) tooling requirements,
(5) time standards (work study),
(6) plant layout (see Section 15.8), and
(7) assembly layout (see Section 15.8).

In other types of industry not all of these planning stages are needed. It will be noted that in explosive industries, production planning is an on-going task, requiring a wide range of new decisions every time a product is modified, or a new product is introduced. In the process industries, however, there is a low rate of design change, and production planning is mainly concerned with pre-planning, or with the initial design of the plant and the plant layout for the factory. The others fit somewhere between these two extremes.

The main controls associated with production planning are process control and maintenance.

15.7.1 The planning specification

The decisions made in production planning depend, in part, on the first level strategic plans made for production. One must know, for example, what quantities of the product are likely to be required per year and the expected life of the product. This information takes the form of a long-term (several years ahead) sales programme.

Next, one needs to know the schedule for product introduction. In other words, when must the first product be produced, and how much time is available for planning and for obtaining any necessary materials and tools?

One also needs a general understanding of production control policy in the company. If, for example, the company is satisfied with an existing multicycle ordering system (stock control or MRP), then order frequencies will be low, and the main emphasis in production planning will be on the minimisation of operation times. If, however, the company needs to reduce stocks and is moving towards just-in-time (JIT) production control, then methods should be chosen which reduce setting-up and throughput times, and are more suitable for use with high order frequencies.

One also needs to know the 'sanction quantity' for the product. For how many products, in other words, do the directors authorise their managers to commit the company, by placing contracts with suppliers for future deliveries of materials and bought components?

Again, the planners must know what is the tooling 'write off quantity' for the product. In other words, what is the number of products over which the cost of any new tooling must be recovered in the selling price? The write-off quantity will affect the choice of selling price for the product and/or the possible investment in tooling. Finally, one needs to know how much money is available for investment in new machines and other processing facilities.

It is best if these decisions are written together in a 'planning specification', which clearly tells the production planners what criteria they are to use when choosing between alternatives during production planning. An example is given in *Figure 15.13*.

15.7.2 Make or buy

The first task in production planning is to go through the product parts lists and decide for each part and subassembly whether provision should be made to produce it in the factory or if it should be bought.

In its wider context, this decision is concerned with choosing whether to provide processing facilities for making castings, forgings, or gears for example. Today, with group technology, these questions can be rephrased to ask: Are we going to provide groups for gear manufacture, for heavy beds and cases, for small rotational parts from bar, for sheet metal pressings, for welded fabrications and others?

The main criteria for choice in this make-or-buy decision-making area are:

(1) economic,
(2) security, and
(3) flexibility.

The problem of the economic choice between make or buy is discussed in Section 15.12. It is much more complex than a simple comparison of manufacturing costs and purchase price.

Planning specification	
PRODUCT Blenker	SPECIFICATION NO. 207.30117
A SALES PROGRAMME 1969 / 1970 / 1971 / 1972 / 1973 50 / 1200 / 1500 / 1750 / 2000	
B INTRODUCTION PROGRAMME 1. Prototype. *Already tested* 2. Pre-production batch. *150 sets already produced* 3. Production tooling. *Complete by 1.10.69* 4. Start prod. assembly. *1.12.69*	
C CYCLE 1. Standard. *2 weeks. (Cat. A and B made. Cat. A bought)* 2. Cat 'B' bought. *8 weeks* 3. Cat 'C'. *10 weeks*	
D 'SANCTION QUANTITY' 1. Up to 1250 products. *1250 sets* 2. After 1250 sets. *3250 sets (confirm Jan 1 1970)*	
E TOOLING BUDGET 1. Special tooling. *£45 000* 2. Write off quantity. *4 500 (end of 1972)*	
F TARGET PRODUCTION COST 1. Direct materials. £15.50 2. Direct labour. £ 7.02½ 3. Direct expense. £ 6.08 Total direct cost £28.60½	
Issued by. *J Smith*	Date. *1.1.69*

Figure 15.13 A planning specification

The fear that others may learn how to make the product and become competitors may persuade some companies that they need to make the more complex parts of their products themselves, for reasons of security.

Finally, the trend in most product markets today is towards an increase in the rate of design change. This makes flexibility—or the ability to change product designs rapidly with minimum losses due to obsolescence, making it possible rapidly to follow changes in market demand—a subject of growing importance. Companies which can make most of the parts they need are generally more flexible in this sense than are those who buy a large proportion of the parts they use.

Make-or-buy policies found in practice include: make all the parts needed; buy all the parts needed; and provide facilities so that most of the different parts needed can be made, but only provide full capacity for the security items, subcontracting some of the other items in order to achieve the necessary capacity.

15.7.3 The choice of material form

There is no problem in choosing the form of material in the process and implosive industries, because in both cases production starts with a small number of material varieties. There is little or no choice between alternative material forms. The square industries are mainly composed of service factories which subcontract for other industries. In this case the choice of material form is generally made by the customer.

It is only in the explosive industries that there is a choice. In essence, we need to decide on the point at which we wish to start processing. A steel part, for example, may start as a cast-steel billet, be rolled into bars, be machined, be electroplated, and then be assembled with other parts to make a finished mechanical product. In this case we need to decide:

(1) Do we want to cast billets?, or
(2) Do we want to buy billets and roll bars?, or
(3) Do we want to buy bars and machine parts?, or
(4) Do we want to buy machined parts, and then electroplate them?, or
(5) Do we want to buy electroplated finished parts, and then assemble them?

If we start at the beginning, we can hope for maximum flexibility because a small variety of materials can be used to produce a very large variety of parts. Because in this case materials are bought in their lowest form, material costs will be low, but the investment in plant and tooling will be high, as will be the expenditure on direct labour. If, on the other hand, we consider buying components in a form which is ready for assembly, we need to know a great deal more about the relative investments and expenditures before we can make a reasonable choice.

At each stage of processing a part, there will also be a choice. For example, if the first stage is to produce a casting ready for machining, one needs to choose between using sand castings, the Cosworth method, die castings, and the lost wax process, among others. Again, to make a gear one might start with a blank cut from bar, or pay more for a hot pressing, which would improve grain flow and quality, reduce the material waste per part, and also, if made in the factory, would help to reduce bought material costs.

15.7.4 The choice of processing route

The next problem is to choose the processing route for each component. A typical 'route card' for an engineering product is illustrated in *Figure 15.14*. The figure shows the 'process' of machining a particular part, divided into 'operations' which are listed and numbered according to their sequence of application. Details are given of the machine to be used for each operation (in *Figure 15.14* by means of a two-letter code) and the standard operation time.

Some companies also give details of machine set-up times. This information is needed for scheduling, but is of doubtful use for loading, because the set-up time varies greatly with the sequence in which different parts are processed. For loading it is more accurate to compare the net load (or the sum of the operation times) with the net capacity, or capacity available for useful work after correcting for set-up time and other losses.

The division into operations is made according to rules, such as the following.

(1) If following operations are carried out on different machines with different operators, they should have different operation names and numbers.

KL11-7 DERRICK CONTROL FAB. 1 9 BRACKET M34099			PRN 4546
FABRICATION 31/ 41/ 40/ 40/ 31/ 71/ 982 503 434 434 982 987 VR W MV DM DM W SA STO			
10 W MV 31/982	Assemble and weld complete, except items B. *Note*: Tack F to B before loading to fixture P6607.	10 00 54	
20 MV W DM 41/503	Load to table and mill feet to 3/8 in. diameter.	30 00 17	
30 DM MV DM 40/434	Load to table pit jig and drill four 17/32 in. diameter holes in feet J4749.	20 00 06	
40 DM DM W	Load to jig drill and ream 1.5005 in./1.4995 in. diameter holes J4748.	60 00 43 P.P.	
50 W DM SA	Stud weld three studs. Part 'H' TM6606	— 00 09$\frac{1}{2}$	
60 SA W VR 71/987 STO	Deburr including all holes. Final inspection.	00 00 04	

Figure 15.14 A route card

(2) If one man operates two or three machines as a work centre, only one operation number is needed for the work done at the work centre.
(3) If a series of operations is carried out on one machine, but the first of these operations must be completed for the whole batch and the machine must then be re-set for the next operation, then each operation should have a different operation name and number.

There is often a wide choice possible in the number of operations per part. Reducing the number of operations, reduces waiting (queuing) time, and also reduces throughput times and stocks. The present trend is, therefore, towards a reduction in the number of operations.

The number of operations per part can be reduced by such methods as:

(1) Planning one complex operation on a machine, to take the place of several simple operations.
(2) Moving a succession of operations on simple machines onto one more complex machine, e.g. changing several drilling operations on single-spindle drills into one operation on a coordinate drilling machine, or changing several operations on milling, boring and drilling machines into one operation on a numerical control machining centre.
(3) Adding short manual operations, such as deburring and inspections, to machine operations with long operation times, so that they can be done by the machine operator between loading and unloading in the machining cycle.

15.7.5 Planning tooling requirements

Having decided on the route for a part, the next job is to plan the tooling, which must be provided at the workplace in order to carry out the operations done there. Tooling can be classified into three main types:

(1) fixtures,
(2) consumable tools, and
(3) jigs.

Fixtures hold materials or assemblies in a convenient position for work to be done on them. Consumable tools are items such as drills, reamers, turning tools, milling cutters, grinding wheels and files, which wear due to direct contact with materials, and have to be replaced at intervals. Jigs are like fixtures in that they hold materials, but they also guide tools so that they operate in the correct positions on the held materials. A typical example is a drilling jig which holds the materials to be drilled, and also guides drills by means of drilling bushes, so that holes are drilled in required positions.

In designing tooling, there are four main objectives:

(1) *quality*—the tooling must be sufficiently rigid and wear resistant that it is possible to produce consistent results;
(2) *speed*—the tooling must be sufficiently rigid and vibration free to work consistently at optimum speeds;
(3) *set-up time*—it must be possible to change the tooling set-up very quickly for one part after another; and
(4) *cost*—inside the limits set by the first three objectives, the cost of the tooling should be minimised.

A simple example of an assembly or manual processing fixture which follows these objectives is a vice with interchangeable soft jaws. Most fixtures for these purposes can follow the same principles of a basic fixture, with adaptive attachments to convert it for work on other similar components. A similar approach can also be used for machining fixtures. Examples are a collet chuck with interchangeable collets and an adjustable jaw chuck with interchangeable jaws.

In consumable tooling, the trend in the case of cutting tools has been towards small interchangeable tool 'bits'—often of hard tungsten carbide or ceramic materials—in accurately machined tooling holders. Tool bits designed in this way can be changed rapidly when they wear, and the tools can be pre-set in their holders for rapid set-up changes.

In the case of jigs, there is a trend towards a reduced demand for this type of tooling. The introduction of coordinate drilling machines and of numerical control machining centres, has for example, reduced the need for drill jigs. Numerical control machines also simplify, or eliminate the need for jigs, with other processes.

15.7.6 Work study

Work study is a technique designed for studying the methods used in industry to do work. It uses two subtechniques: method study and time study. Method study analyses existing or planned methods and, by eliminating unnecessary or wasted movements, attempts to find the simplest and most efficient way to complete a task. Time study measures the time required to complete tasks in order to provide the necessary data for scheduling and/or loading. Work study was originally developed to study work done in manufacturing workshops. A very similar technique called 'organisation and methods' (O&M) is used to study the work done in offices.

Method study generally starts by breaking down the work required to complete an operation, into steps. The steps are listed in sequence and classified to show for each step if it is: work, transportation, delay, storage or inspection. Transportation distances and work step times, are estimated and added to the chart. An attempt is then made to redesign the method to eliminate the transportation, storage and delay steps. An example is given in *Figure 15.15*.

A more advanced form of method study is called 'motion study'. This was developed in the USA by the late Frank Gilbreth and his wife. They broke down tasks into a series of fundamental elements which they called Therbligs (Gilbreth spelt backwards) see (*Figure 15.16*), and then rebuilt an ideal cycle from which all unnecessary elements were eliminated.

'Time study' was originally based on the time taken to do a task, as measured with a stop watch. The method was used as follows.

(1) *Preliminary study*—make sure that the correct method is being used.
(2) *Divide into elements*—the elements must be easily identified and measured.
(3) *Time the element*—take a number of readings of the time required per element.
(4) *Find the 'recorded element time'*—strike out abnormal times and use the average, minimum, modal or good time method, to find the recorded element time.
(5) *Levelling or performance rating*—adjust the recorded element times to the level that can be expected from an average man or woman.
(6) *Allowances*—add time allowances to cover delays, rest and personal needs.
(7) *Standard time per operation*—specify.

Time study with stop watches is not usually liked by the workers in factories. Synthetic times can be used to avoid the need to use a stop watch. With these methods, standard operation times are found by adding recorded standard time values for standard 'basic work elements' or motions.

Another method used for measuring operation times in small batch and jobbing production is known as 'slotting'. A list is made of the operations done on a machine. These are listed in their estimated operation time sequence. Time study is used to find the operation times for selected operations in the range (bench marks). New work is slotted into the series between bench marks (i.e. more time consuming than this task; less time consuming than the next one). An operation time is allocated to the new job which is half way between those for the adjacent bench marks.

15.8 Design of the material flow system

The material flow system in a manufacturing company is the system of routes along which materials flow between the places where work is done on them in order to change their shape, form or nature. Its nearest equivalent in everyday life is a detailed road map. There is evidence that the efficiency of a manufacturing system depends on the complexity of its material flow system. Simple flow systems are generally much more efficient than are complex flow systems. It is also readily apparent that the nature of the material flow system depends on the way in which the factory is organised at the level of the workshop. This can be taken as our point of departure for an examination of material flow systems.

15.8.1 Organisation at workshop level

Organisation has been defined as: 'the process of determining the necessary activities and positions within an enterprise, department or group, arranging them into the best functional relationships, clearly defining the authority, responsibilities and duties of each, and assigning them to individuals so that the available effort can be effectively and systematically applied and coordinated.

There are two main types of organisation used at workshop level:

(1) process organisation, and
(2) product organisation.

With process organisation, organisational units specialise in particular processes, e.g. turning, milling, grinding, drilling, assembly, heat treatment, painting, and press work. The people in such units specialise in the skills required to do each process.

With product organisation, organisational units complete particular products, assemblies, major stages in assembly, parts, or 'families' of similar parts. They are equipped with the variety of machines and other manufacturing facilities needed to complete such entities, and the people in these units have between them all the skills needed to do so.

There are two main types of product organisation found in industry:

(1) continuous line flow, and
(2) group technology.

With continuous line flow, the machines and/or other processing facilities used to make parts, are laid out in a line one after the other in the sequence in which they are used. The materials being processed flow between these workstations in a continuous stream.

With group technology, products or components are completed in their own special organisational units (or groups). In the case of components, different items may use different combinations of facilities in different sequences inside the groups. In the case of assembly groups, products may be completed by single groups, or by sets of groups working together. In both cases the groups produce continuously, and

Product: *B. B. Haymaker*	Date: *2.1.78*
Operation: *Riveting flier arm assembly (AA200)*	Study by: *J. Smith*
Department: *Sub-assembly*	Sheet No.: *1 of 1*

Symbols: ○ Work ⇨ Transport D Delay ▽ Storage □ Inspect

Time	Travel in m	Symbol	Step No.	Description
				Present Method
25 sec		○	*1*	*Press pin (AA213) into bracket (AA215) on Hyton press (2073). Place in box pallette*
	15	⇨	*2*	*Transport to work in progress store*
		▽	*3*	*Store*
	15	⇨	*4*	*Transport to Hyton press (2073)*
15 sec		○	*5*	*Squeeze rivet end of pin. Place in pallette*
	15	⇨	*6*	*Transport to work in progress store*
		▽	*7*	*Store*
	15	⇨	*8*	*Transport to Hyton press (2073)*
30 sec		○	*9*	*Rivet arm (AA219) to bracket and pin assembly (AA210). Place in pallette*
	35	⇨	*10*	*Transport to welding department*
				Proposed Method
50 sec		○	*1*	*Press pin (AA213) into bracket (AA215) and rivet, then rivet on arm (AA219). New fixture for Hyton press. Place in pallette*
	35	⇨	*2*	*Transport to welding department*

Note. In addition to savings in transportation and storage, estimate operation time saving of 20 seconds each.

Summary:	○ Steps	⇨ Steps	⇨ Metres	D Steps	▽ Steps	□ Steps
Present Method	*3*	*5*	*95*		*2*	
Proposed Method	*1*	*1*	*35*		*0*	*Q.C.*
Saving	*2*	*4*	*60*		*2*	

Figure 15.15 An operation analysis sheet

make most of their own subassemblies. With group assembly, a main objective is minimum throughput time and sub-assemblies are not usually made in batches for stock.

15.8.2 Continuous line flow

Continuous line flow is used in practice in: process industries (see *Table 15.1*); in the mass production of products in very large numbers; and for continuous manual operations, such as assembly and packaging, for segments of material flow systems where the production facilities are normally always used in the same sequence, and the operation times are similar in length.

Continuous line flow infers a need to balance the operation times at each station on the line. If this is not done, either stock will build up between stations where a long operation follows a short one, or there will be idle times when some operations must wait for materials at each cycle, from the previous operation. The automated version of continuous line flow is the automatic transfer line.

The ideal situation for continuous line flow is found in the process industries. These convert a small number of material varieties into an equally small number of product varieties (see *Figure 15.2*). Conversion is achieved using a small number of processes, generally in the same sequence. The only limitation to the use of continuous line flow in process industries is found

Therblig	*Abbreviation*	*Symbol*	*Suggested by:*
Search	Sh		Eye searching
Find	F		Eye straight
Select	St		Reaching for object
Grasp	G		Hand open to grasp
Transport loaded	TL		Hand with item in it
Position	P		Hand placing object
Assemble	A		Four parts together
Use	U		Word 'use'
Disassemble	DA		Assembly one part removed
Inspect	I		Magnifying lens
Pre-position	PP		A nine-pin set up for game
Release load	RL		Hand dropping item
Transport empty	TE		Empty hand
Rest	R		Man seated
Unavoidable delay	UD		Man falling down
Avoidable delay	AD		Man lying down
Plan	Pn		Finger to brow (thinking)
Hold	H		Magnet and iron bar

Figure 15.16 Therbligs

in those cases where there are some batch processes. The long-term expectation is that most of these batch processes can be converted to continuous flow. For the time being, they require intermediate batch processing stages in otherwise line-flow factories.

The use of continuous line flow for the mass production of components for sale or to feed assembly is diminishing due to the demands of markets for a greater variety of product designs. This trend is likely to continue, leading to a gradual reduction in the number of automatic transfer lines for component processing, and an increase in the use of flexible manufacturing systems (FMSs).

Continuous line flow in assembly can be found in the forms of machine-paced lines (mechanical conveyors) and operator-paced lines. The future use of machine-paced continuous line flow for manual assembly is likely to diminish, (a) because such lines are generally unpopular with the people who work on them; (b) due to the present explosion in the number of product variants, which reduces the efficiency of such lines; and (c) due to the gradual development of mechanised and automated assembly.

Simple assembly lines comprising a few stations making simple assemblies do have a future. In this case, balance is achieved by a team of workers who move between stations as necessary to maintain the flow of work. This type of layout is often used within assembly groups.

Where lines must be used, they have the advantage of short throughput times, leading potentially to low stocks and low stock-holding costs.

15.8.3 Group technology—component processing

The term 'group technology' (GT) was first used by Professor S. P. Mitrovanof of Leningrad University, as the name for his research into the relationship between the shape of components and the processing methods used to make them. One of his findings was that it was possible to set up lathes so that they could make a set of similar parts, one after the other at the same set up. Later, a company in Alsace added milling and drilling machines to a section containing lathes set up in the Mitrovanof manner, to form a group which completed all the parts it made. Later again, in Britain, it was shown that this method of organisation into groups which completed all the parts they make, could also be used with other processes, and in other types of industry.

At this stage group technology had evolved into a method of organisation which could be used in any factory where continuous line flow was not possible. Group technology makes it possible to gain the advantages of continuous line flow, with jobbing and batch production, in implosive, square, and explosive industries.

The difference between process organisation and group technology is illustrated diagrammatically in *Figure 15.17*. With process organisation, materials move from one processing unit and foreman to another during manufacture, until all the manufacturing processes have been completed. In each unit they join the queue of work waiting to be processed and must then wait for their turn to be processed. This type of organisation is typified by very long throughput times. It is not

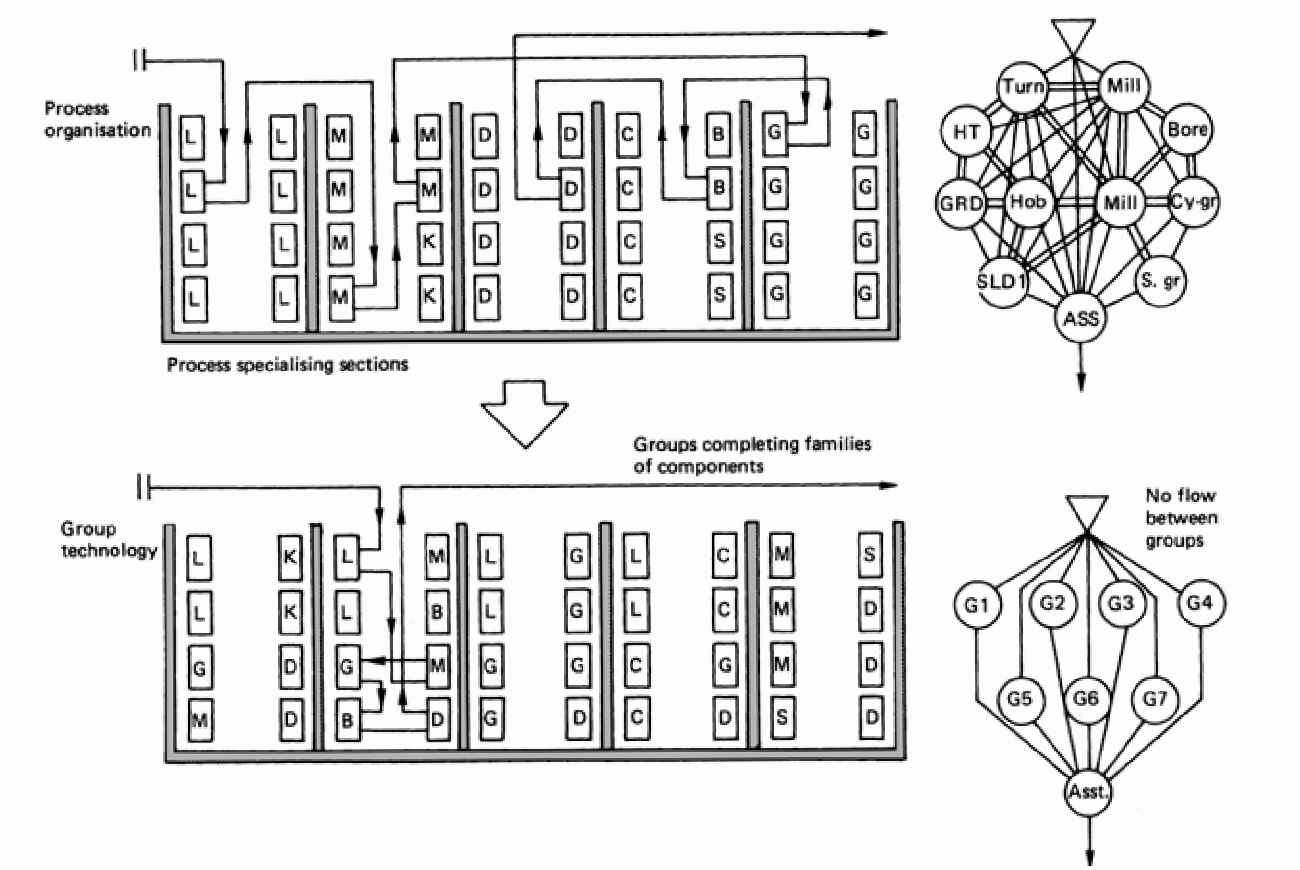


Figure 15.17 The difference between process organisation and group technology. Group technology is a form of product organisation

uncommon to find that the throughput time in a processing department with process organisation is over 100 times greater than the sum of the operation times. With group technology, each part is completed in its own special group under the supervision of its own foreman. In these circumstances it is possible to schedule the work to achieve very short throughput times.

Figure 15.18 shows a group installed about 25 years ago. It was designed to produce simple rotational parts made on bar lathes. This group completed all the parts it made, and was responsible for its own inspection, operation scheduling, setting-up, tool storage, minor tool maintenance, preventive maintenance inspections up to the minor inspection of machine tools, and for its own housekeeping. The group was, in effect, a semi-independent mini-factory.

The group shown in *Figure 15.18* was equipped with manually operated bar capstan lathes, mills and drilling machines, and with barrel deburring machines and cleaning tanks. Today, the capstan lathes have been replaced by CNC turning machines, one of which is capable of both first and second operation turning. Much of the milling and turning is done by power tooling on the turning centres, but a small machining centre and a drilling machine have also been installed.

Many of the product designs have changed, but there is still a need for small rotational parts. The present group produces a greater volume of work than the old one, has half as many machines, and employs half as many people. It provides an example of the gradual evolution of groups towards automation.

Experience has shown that, providing one starts with a department which completes all the parts it makes, and which employs, say, 20 or more people, one can, with very few exceptions, divide it totally into groups which complete all the parts they make. The exceptions are generally processes needed by more than one group, of which X-ray and heat treatment are examples from practice. Where such processes cannot be divided for assignment to different groups, they are organised as a service group, which can provide a service to any group that needs it. Service groups should always be treated as undesirable exceptions which need to be eliminated as soon as possible.

The number of people in a group depends partly on the complexity of the product. If some complete parts are made with operations on 40 different machine tools, then there must be one group with at least 40 machines installed, with enough people to man them. Experience has shown that with group sizes in the range 10–15 people per group, with one or two groups in some factories employing a few more people for complex parts, then it is generally possible to divide into groups without buying any new machines.

Table 15.2 lists the main advantages of group technology. Most of these advantages arise because groups complete parts and their machines are close together under one foreman. The

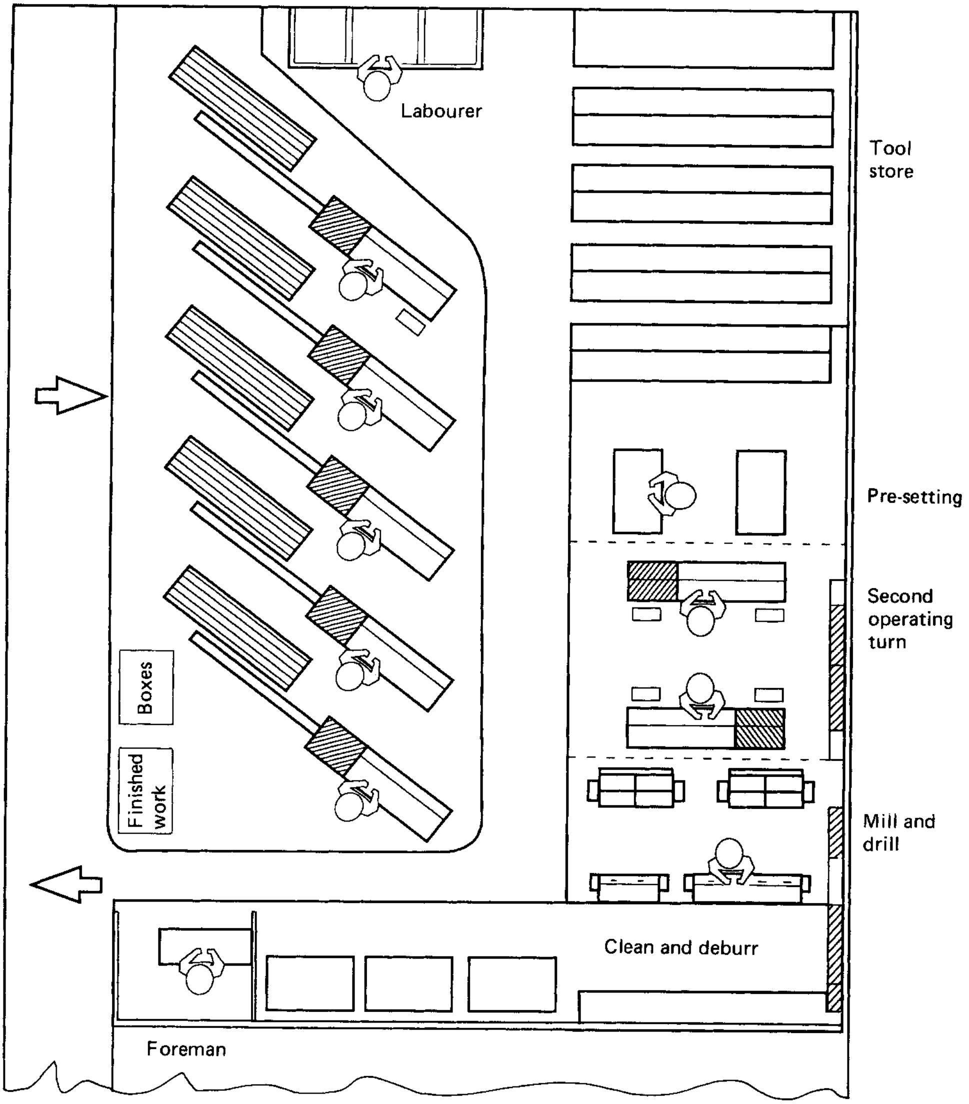


Figure 15.18 A group technology group. The group carries out its own operation scheduling, inspection, materials handling, setting up, tool storage and minor repairs, minor plant inspections and maintenance, and housekeeping

maximum advantages are only obtained, however, if the change to group technology is total and company wide. Group technology is not a technique which can only be used for some parts and not for others.

Figure 15.19 shows the effect of group technology on the material flow system of an enterprise. *Figure 15.19(a)* shows a network diagram which illustrates the flow system with process organisation. The circles represent processes and the arrows between them show the routes followed by materials between processes. This illustration is not some monstrosity chosen for effect, it is typical of most companies in the mechanical engineering industry that use process organisation.

Figure 15.19(b) shows the same company but with group technology. There is flow between groups at different processing stages, but because with very few exceptions the groups complete all the parts they make, there is no flow between groups at the same level or processing stage. It can be seen that group technology greatly simplifies the material flow system of an enterprise.

15.8.4 Classification and coding

The main difficulty in introducing group technology into industry has been in finding a way to plan the change from process organisation to group technology. Most of the early group technology applications were planned, using classification and coding (CC). This method classifies the information contained in part drawings, and is based on the idea that parts which are similar in shape can normally be machined by the same group of machines.

Table 15.2 Advantages of group technology

1. Short throughput times
Because machines are *close together* under one foreman, giving:
(a) low stocks,
(b) low stock holding costs, and
(c) better customer service

2. Better quality (fewer rejects)
Because groups *complete parts*, and machines are *close together under one foreman*

3. Lower materials handling costs
Because machines are *close together under one foreman*

4. Better delegation
Because groups *complete parts*. The foreman can be made responsible for cost, quality and completion by due-date giving:
(a) reduced indirect labour cost, and
(b) more reliable production

5. Training for promotion
Process organisation only produces specialists. Group technology trains generalists

6. Automation
Group technology is the first evolutionary step in automation. A group is a flexible manufacturing system with some manual operations

7. Morale & job satisfaction
Most workers prefer to work in groups than with process organisation

Although this idea is true for the majority of parts, it is not a reliable method for finding groups, because there are too many exceptional cases when it is untrue. For example, parts may be similar in shape but vary in size, manufacturing tolerances, requirement quantities, or materials, making it necessary to manufacture them on different sets of machines.

A second disadvantage of CC is that it fails to find the large numbers of parts in most factories which differ in shape or function from others, but ought to be made in the same groups because they are made on the same machines. For this reason, CC never finds a complete division of the parts made in a factory into families.

A third disadvantage of CC is that, because it is based on the analysis of component drawings, it does not find groups (of machines), but only families of parts. After CC has produced a list of parts arranged in 'families' together with a long list of exceptions, there is still a great deal of expensive work to be done in production planning for the groups.

Finally, CC ignores the effect of its decisions on the material flow system and, in particular, it fails to start with a division into departments which complete all the parts they make. Many of the early problems in planning group technology could be traced to the fact that it is impossible to divide a department which does not complete all the parts it makes into groups which do complete all the parts they make.

If one starts, for example, with a 'machine shop' (department) for which 15% of the parts have intermediate operations in some other department, and another 15% have intermediate subcontracted operations, there is no possibility of forming groups until these anomalies have been corrected. There is no mechanism in CC for making this necessary correction.

In an attempt to develop CC so that it can be used to plan group technology groups, some manufacturing and computer software companies have developed data with 60 or more digits per component. In these companies CC has become an esoteric and very expensive computer game. As long as these banks are restricted to the information contained in component drawings and parts lists, such systems are still unable to find an efficient division into group technology groups. Even if production data are added to the component data, they only provide the information needed for later production flow analysis, at unnecessarily high cost.

It should be stated that, although CC is an inefficient method for planning group technology, it can be used efficiently to reduce drawing-office costs (by finding existing designs which can be used again) and to reduce tooling costs (by finding tools for similar parts which can be adapted at a low cost for a new part, instead of making new tools). It is also possible that CC may be needed for the future development of computer-aided design and manufacturing.

15.8.5 Production flow analysis

An alternative method to CC for planning group technology groups and families is production flow analysis (PFA). Where CC finds 'families' of parts by analysing the information in component drawings, PFA finds both 'groups' of machines and 'families' of parts, by analysing the information in component route cards.

Where CC is based on the idea that parts which are similar in shape or function can be made by the same set of machines, PFA is more direct. PFA considers that parts that are made on the same set of machines should be made together in a group which contains this set of machines. Because most parts which are similar in shape are made on the same machines, PFA (like CC) tends to bring such parts together into the same groups, unless they vary greatly in materials, size, tolerance or requirement quantities, and are made for these reasons on different machines. PFA groups also include parts which are not similar in shape or function if they are made on the same machines, and for this reason is able to find a complete division of all made parts into families.

PFA uses five subtechniques one after the other to plan group technology. Four of these techniques are illustrated in *Figure 15.20*. Company flow analysis (CFA) is used to simplify the flow of materials between the different factories, or major product divisions, in a large company. Factory flow analysis (FFA) plans a simple material flow system between the departments in a factory. It does this by designing departments which, with very few exceptions, complete all the parts they make. Group analysis (GA) divides each department into groups which complete all the parts they make. Line analysis (LA) analyses the flow of materials between the machines in each group, to help in finding the best arrangement for group layout. Tooling analysis (TA) analyses the tooling used on each machine in each group, in order to rationalise the tooling and to find 'tooling families' of parts which can be made at the same set-up.

15.8.6 Factory flow analysis

In most manufacturing enterprises PFA starts with factory flow analysis (FFA). The first need is to record the existing material flow system as illustrated in the simple example given in *Figure 15.21*. The processes used in the factory are given a

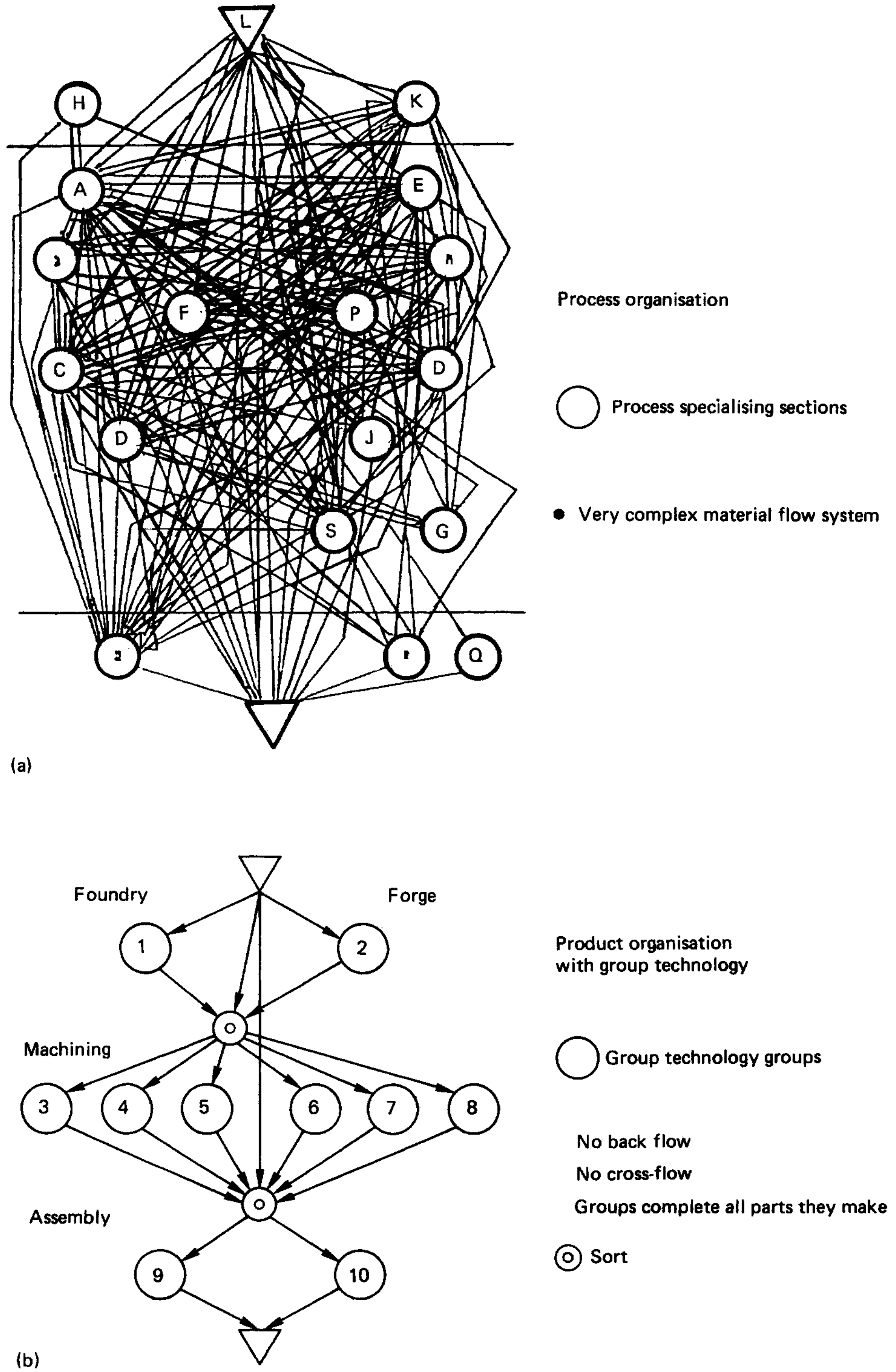


Figure 15.19 The change engendered by changing from process to product organisation. It can be seen that the use of group technology greatly simplifies the material flow system

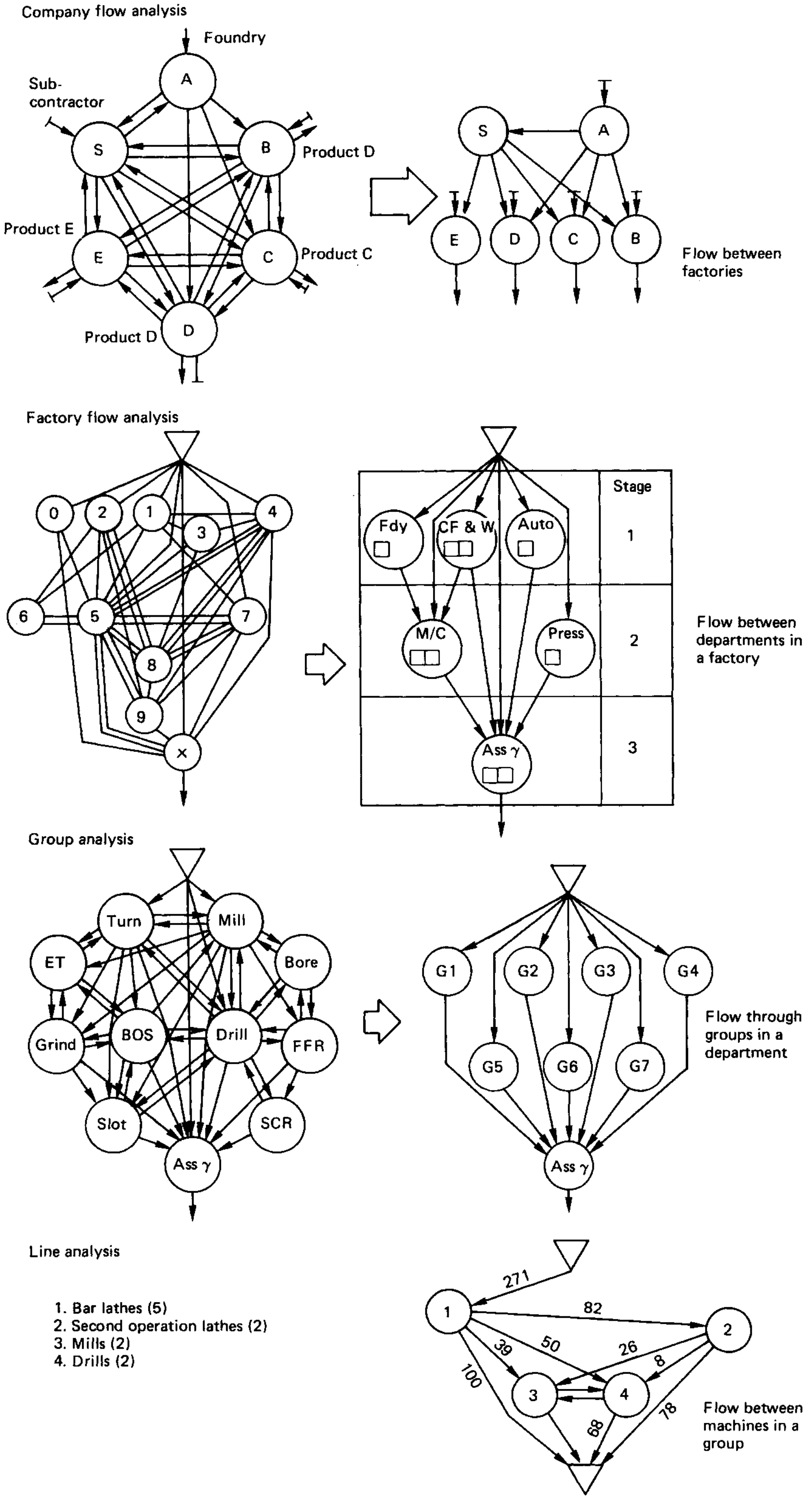


Figure 15.20 Subtechniques of production flow analysis

Process code

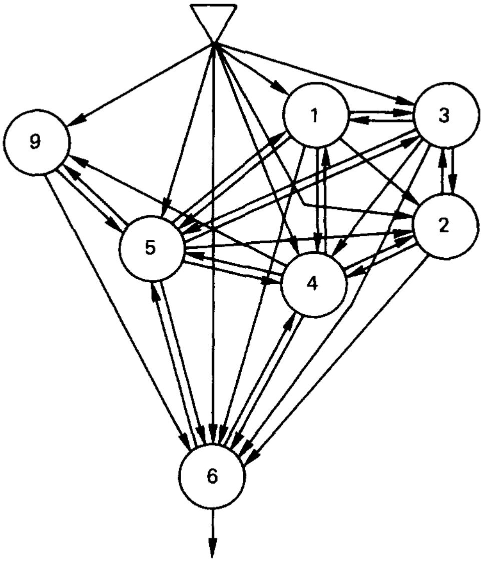
Code	*Process*
1	Blank production
2	Sheet metal
3	Forging
4	Welding
5	Machining
6	Assembly
9	Subcontracting

From/to chart

	To									
From	*1*	*2*	*3*	*4*	*5*	*6*	*9*	*From*	*Ends*	*Sum*
1	—	10	27	133	57	1	—	228	—	228
2	—	—	1	1	—	43	—	45	—	45
3	1	3	—	5	5	32	—	46	—	46
4	3	1	—	—	8	50	1	63	145	208
5	1	3	3	10	—	80	3	200	—	200
6	—	—	—	1	1	—	—	2	365	367
9	—	—	—	—	3	2	—	5	—	5
Into	5	17	31	50	74	308	4	—	510	1099
Start	223	28	15	58	126	59	1	510	—	—
Sum	228	45	46	208	200	367	5	1099	—	—

PRN frequency chart

PRN	*No.*	*PRN*	*No.*	*PRN*	*No.*
1 2 3 4	1	1 5 4	3	4 1 4 6	2
1 2 6	8	1 5 4 6	1	4 1 5 6	1
1 3 2 4	1	1 5 4 9 9 5 6	1	4 2 6	1
1 3 2 6	1	1 5 6	142	4 5 4 6	1
1 3 4	3	1 5 6 4	1	4 5 6	7
1 3 5 6	3	1 5 9 5 6	1	4 6	46
1 3 6	18	1 6	1	5 4	4
1 4	131	2 6	28	5 6	120
1 5 1 3 2 6	1	3 1 2 6	1	5 9 5 6	1
1 5 2 6	3	3 4	1	5 9 6	1
1 5 3 5 6	1	3 5 6	1	6	59
1 5 3 6	2	3 6	12	9 6 3 6	1
Total					510



Material flow network

Figure 15.21 Factory flow analysis record of an existing flow system. PRN, process route number

code number. Each part is then given a process route number (PRN) showing the code numbers of the processes used to make it in their sequence of use. A PRN frequency chart and a from/to chart are produced and this information is used to draw a network diagram of the existing material flow system.

Simplification of this system starts by drawing the primary-flow system, which is based on the PRNs which include the largest number of parts. As shown in *Figure 15.22*, this primary-flow system includes a large proportion of the made parts, but uses a very small proportion of the different PRNs. It is much simpler than the original network based on all the PRNs.

The primary network is now simplified by:

(1) combination, and
(2) redeployment of plant.

Combination brings together closely related processes to form departments. As shown in *Figure 15.23*, this combination makes it possible for many other PRNs and their parts to join the simplified system.

Redeployment of plant then makes it possible for these new departments to complete all the parts they make. As an example, a cold forging and welding department (see *Table 15.3*) sent 24% of the components it made to a machining department for intermediate drilling operations. Moving one radial drill to department 34 made it possible for them to complete all the components they produced. In other factories, boring machines and milling machines for plate edge milling have also been transferred to cold forging and welding departments.

It is not difficult with FFA to convert a traditional department based on process organisation to a group technology department which completes all the parts it makes. Nearly all the machines in each department will be the same after the change as they were before, and the number of exceptional parts requiring replanning seldom exceeds 1%.

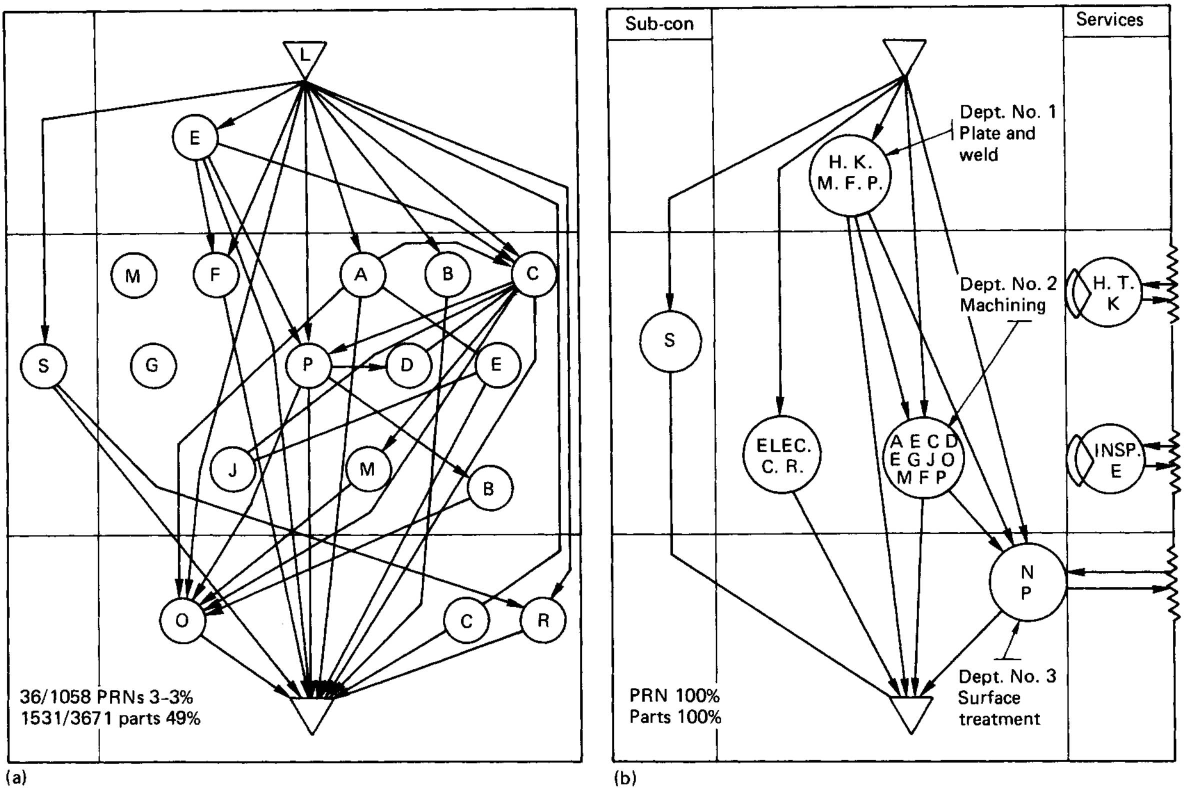


Figure 15.22 Factory flow analysis: (a) the primary-flow system; (b) a simplified flow system

15.8.7 Group analysis

Group analysis is the primary subtechnique of PFA. Providing one starts with a processing department which completes all the parts it makes, and providing this department employs, say, 20 people or more, it is always possible to divide it into smaller groups which complete parts.

As illustrated in *Figure 15.23*, group analysis can be seen as a problem of the resolution of a matrix. *Figure 15.23(a)* is a matrix with parts listed across the top in part number sequence, and machines listed vertically in machine number sequence. Ticks show which machines are used to make each part. It is always possible in such a case to rearrange the sequence in which the parts are listed, and to reorder the listing sequence for the machines in such a way that groups and families are formed. There will usually be some machines of which there are more than one machine of the type, which are needed in more than one group, and there will usually be some exceptional operations which do not fit. In general, however, the division into groups is a natural division which is not difficult to find.

The main problem with the matrix is usually its size. In a company with, say, 4000 current parts in production and 300 machine tools, the matrix is much too big for manual resolution. Even with the computer, it helps if one can divide the machines into sets (types) which can all do any operation on any machine in the set, and can divide the parts into sets (modules), for which there is a high probability that the parts in any set will all fit happily into one group family. These modules are, in effect, mini-groups.

In the case of the machine tools, the ideal division into types would be such that any operation allocated to one machine of a type, could equally well be allocated to any other machine of the same type. In practice it is seldom possible to change the exsting plant code numbers to follow such an ideal, due to the thousands of changes which would be needed in the plant list, route cards and plant layout records. The analyst should know what would be the ideal arrangement, but cannot generally hope to use it.

15.8.8 Group analysis—the division into modules

In the case of the parts, the total matrix (see *Figure 15.23*) can be divided without great difficulty into a much smaller number of submatrices, or modules. Each module is based on one machine—the *key* machine—from the plant list, and contains all the remaining parts which have not been included in previous modules which have operations on the key machine.

Figure 15.24 illustrates a typical module. It consists essentially of a matrix showing the parts in the module, and the key machine together will all other machine types used to make them. For ease of presentation on the computer, additional details about the machines and the parts are generally shown separately, in addition to the matrix.

Table 15.3 A home-and-away chart

PRN	No.	Home dept.	Away Dept	Away Operation	Transfer	Revised PRN
1234	1	34	2	Press	Press to 34 from 2	134
1324	1	34	2	Press	Press to 34 from 2	134
1326	1	34	2	Press	Press to 34 from 2	136
1356	3	34	5	Drill	Drill to 34 from 5	136
151326	1	1	5	Drill	Drill to 1 from 5	136
		34	2	Press	Press to 34 from 2	—
1526	3	5	2	Press	Exception	—
15356	1	34	5	Drill	Drill to 34 from 5	136
1536	2	34	5	Drill	Drill to 34 from 5	136
154	3	34	5	Drill	Drill to 34 from 5	14
1546	1	34	5	Drill	Drill to 34 from 5	146
1549956	1	34	5	Drill	Exception	—
		34	9	SpHT	—	—
		34	9	X-ray	—	—
		34	5	Drill	—	—
1564	1	34	5	Drill	Drill to 34 from 5	14
		34	6	Assembly	Bench operation to 34	—
15956	1	5	9	Grind	Exception	—
3126	1	34	1	Saw	Saw to 34 from 1	36
			2	Press	Press to 34 from 2	—
356	1	34	5	Drill	Drill to 34 from 5	36
4146	2	34	1	Saw	Saw to 34 from 1	46
4156	1	34	1	Saw	Saw to 34 from 1	46
			5	Drill	Drill to 34 from 1	—
426	1	34	2	Press	Press to 34 from 2	46
4546	1	34	5	Drill	Drill to 34 from 5	46
456	7	34	5	Drill	Drill to 34 from 5	46
54	4	34	5	Drill	Drill to 34 from 5	4
5956	1	5	9	Mill	Exception	—
596	1	5	9	Drill	Transfer operation 9 to 5	56
9656	1	5	9	SpHT	Exception	—
			6	Fettle	Bench	—
24	41	—	—	—	—	—

PRN, Process route number. SpHT, Special heat treatment

It will be obvious that some machines—such as drilling machines in machine shops—are going to be needed in most of the groups, and others such as gear hobbing machines will probably be special to one group only. In order to form modules likely to be each in one group only, one must start forming modules based on the special machines. This is done by classifying the machines as follows.

S *Special machines*—there is only one machine of the type. It would be very difficult to transfer an operation on a machine of this type to any other type of machine.

I *Intermediate machines*—these are the same as the S-type machine, but there is more than one.

C *Common machines*—machines likely to be in more than one group, e.g. lathes, mills and drills in machine shops.

G *General machines*—these are widely used machines, of which there are too few to divide between groups, e.g. saw for cutting blanks.

E *Equipment*—this is used to assist manual operations, e.g. vices, surface plates, hand power tools etc.

A special plant list (SPL) is produced, showing the sequence in which the machines in the plant list are to be selected as the key machine. This lists the machines in the sequence SICGE, and in ascending *F*-value sequence within each category, where *F* is the number of parts with operations on each machine.

Figure 15.25 shows a module summary in the form of a matrix containing modules and machines. It is obvious that each module removes at least one machine (the key machine) from the data bank, together with a list of parts. These items are not considered again in forming later modules. The last 22 modules in *Figure 15.25* contain over 50% of the made parts, reflecting the fact that in most mechanical products most of the parts are simple items made on common machines such as lathes, mills and drills.

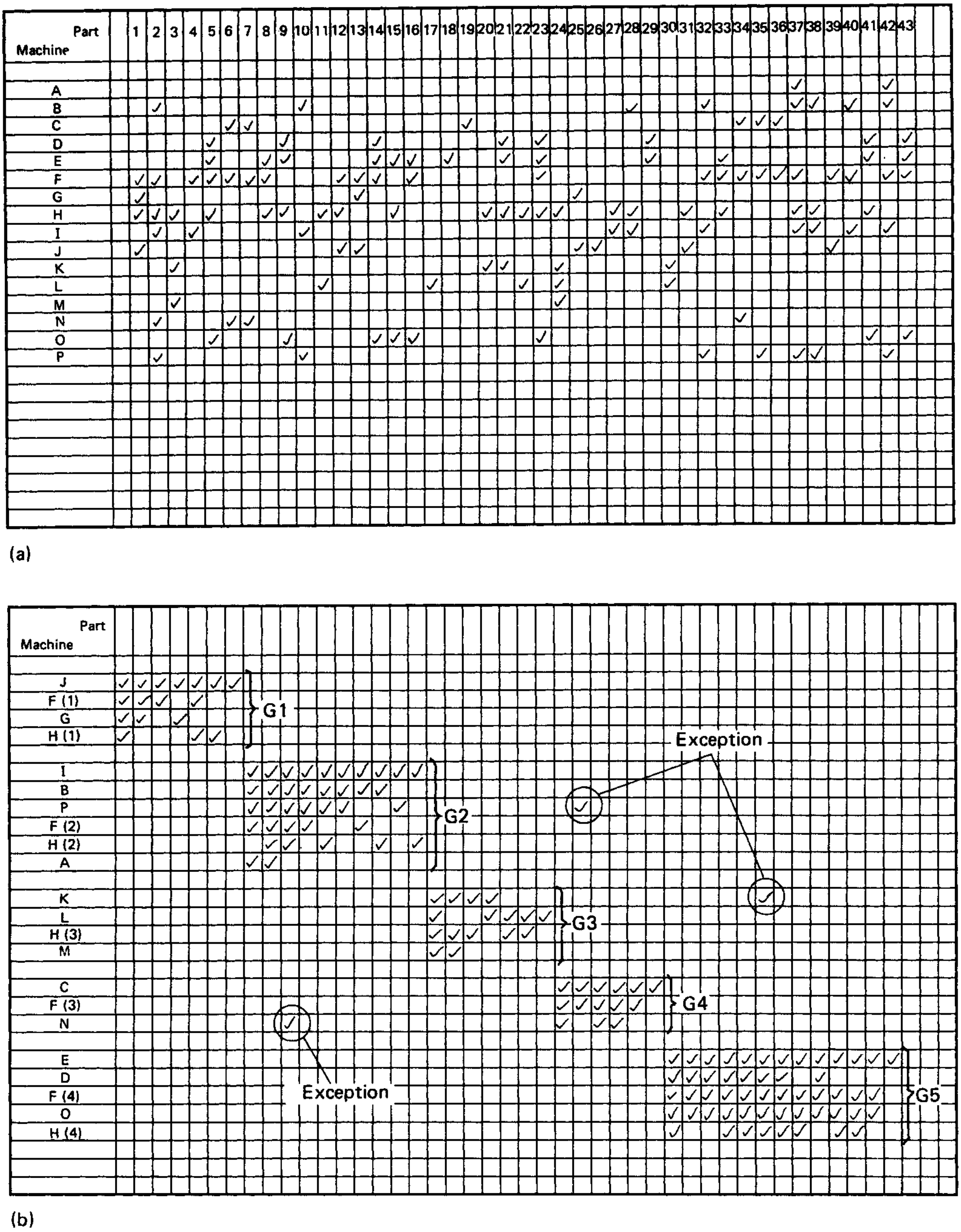


Figure 15.23 Group analysis matrices: (a) component machine chart; (b) division into groups and families

15.8.9 Finding groups and families

Groups are formed by combining modules. The method involves:

(1) choosing likely groups,
(2) checking the modules and module summary for group feasibility,
(3) selecting modules as nuclei for groups,
(4) using the computer to find best fit groups for remaining modules,
(5) allocating machines to groups,
(6) re-allocating exceptional operations,
(7) eliminating remaining exceptions, and
(8) checking the loads on machines allocated to two or more groups.

Machines

Module No.	*Machine No.*	*Machine Name*	*N*	*F*	*SIC GE*	*f*
08	231202	Beoach	1	9	S	9
24	060210	Cyl. grind	1	52	S	1
25	060901	Hone	1	68	S	1
47	040409	Mill vert.	1	46	S	1
26	040607	Engraver	1	92	S	1
30	130104	Geap cut	3	42	I	1
29	060301	Bore grind	2	41	I	1
63	050101	Jig borer	2	112	I	1
31	040606	Pantograph	2	57	I	1
45	010125	Lathe ctr.	1	35	C	1
56	060202	Cyl. grind	4	72	C	1
57	010105	Lathe ctr.	4	80	C	1
58	010123	Lathe ctr.	1	87	C	1
48	040512	Mill univ.	2	47	C	1
68	010119	Lathe ctr.	2	205	C	2
67	050902	Drill coor.	2	146	C	1
65	040510	Mill univ.	2	125	C	1
70	040504	Mill univ.	6	347	C	2
69	010126	Lathe ctr.	7	260	C	2

Matrix

Item	*Part No.*	231 202	060 210	060 901	040 409	040 607	060 301	130 104	040 606	050 101	010 125	040 512	060 202	010 105	010 123	040 510	050 902	010 119	010 126	040 504
1	28947273	⊗							⊗					⊗						
2	28948636	⊗															⊗			⊗
3	28949279	⊗																⊗		⊗
4	28949303	⊗																⊗		
5	28949428	⊗	⊗			⊗									⊗				⊗	
6	38444311	⊗		⊗	⊗			⊗		⊗	⊗	⊗				⊗			⊗	
7	39484005	⊗																		
8	39454624	⊗																		
9	39454731	⊗					⊗						⊗							

Parts

Item	*Part No.*	*Part name*	*Mtl type*	*Mtl form*
1	28947273	End cover	MS	Bar blank
2	28948636	Spacer	MS	Bar blank
3	28949279	End cover	MS	Bar blank
4	28949303	Spacer	MS	Bar blank
5	28949428	Bush	MS	Bar blank
6	38494311	Gear	MS	Forging
7	39454095	Bush	MS	Bar blank
8	39454624	Bearing	MS	Bar blank
9	39454731	Spacer	MS	Bar blank

Figure 15.24 A module

	Machine					Module number and (f)																																																																													
Key	No	Name	N	F	SICGE	1	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83
5	080301	Slotter big	1	1	S		1																																																																														
6	060104	Surf. grind	1	2	S			2																																																																													
7	040906	Borer pre set	1	2	S				2																																																																												
8	040605	BMW brake	1	2	S					2																																																																											
9	010101	Ct lathe long	1	3	S						2																																																																										
10	040603	Copy mill	1	4	S							4																																																																								
11	060102	Surf grind	1	5	S								5																																																																							
12	040604	Copy mill	1	5	S							1		4																																																																						
13	130401	Gear hob	1	6	S										6																																																																					
14	010103	Ct lathe long	1	6	S											6																																																																				
15	060205	Cyl. gr v. big	1	6	S										2		4																																																																			
16	010106	Ct lathe long	1	7	S				1		2							4																																																																		
17	060302	Bore gr shaft	1	7	S													3	4																																																																	
18	060502	Bore gr	1	8	S															8																																																																	
19	080303	Slotter small	1	8	S														1	1	6																																																																
20	108102	Broach	1	9	S										4						5																																																														
21	010303	V. Borer big	1	9	S		1				1												8																																																													
22	231202	Broach	1	9	S																			9																																																												
23	060902	Hone	1	14	S										1									13																																																											
24	010304	V. Borer v. big	1	14	S				1	1																12																																																										
25	060401	Thread grind	1	15	S										1				2							12																																																									
26	040908	Borer v. big	1	16	S		1																			3		11																																																								
27	010702	N.C. Turn ctr	1	17	S						1								1	2	1						1		11																																																							
28	231212	Broach	1	22	S								1						1		1									19																																																							
29	010305	Borer big	1	25	S																					2		2			21																																																						
30	040905	Borer v. big	1	26	S		1			1								2							1				5		2	12																																																				
31	040511	V. Mill small	1	30	S																												30																																																			
32	060206	cvl gr. big	1	30	S					1			1	1	4					2		1					1						19																																																		
33	010705	Turn ctr big	1	35	S									1	4			2	1	1	1	3							2		1			4	15																																																	
34	060103	Surf grinder	1	38	S																		1					1			3			2		31																																																
35	040514	Mill 3 spindle	1	41	S																																41																																															
36	040413	Mill CNC	1	41	S															2					1										4		34																																														
37	040904	Borer v. big	1	68	S		1		2							5				1			3			1		1	3		1	1							49																																														
38	060210	Cyl grinder	1	68	S										1								1			2			1			8	1	1		1			57																																													
39	060901	Hone	1	78	S																1			1			2			1			3	1	1	1	1			17	43																																												
40	040409	Mill vert NC	1	106	S																			1	1				1					3	1	1	1		5	5	87																																										
41	040807	Engraving M/C	1	135	S							1					1							1						1			3							16	16	8	88																																									
42	040909	M/C centre	2	25	I																1				2									2			1		2		1		16																																								
43	010128	Lathe Micron	2	27	I							2					2																2								11		1		9																																							
44	060101	Surf grinder	2	32	I					1										4					2								3	1	2	1			2				1		15																																						
45	130104	Gear cut M/C	3	37	I																			1						1			2							8	84		1																																									
46	040907	Borer CNC	2	46	I		1										3		1	2			1			4		1	4		3	1			1	1			3			6						14																																				
47	060301	Bore grinder	2	52	I					1			1		1					4	1	2		1	1		3		1	1	1	1		9	1						1					3			19																																			
48	050101	Jig borer	2	53	I																2			1						1				1	1	1	1	1		5	1	17	3							18																																			
49	040606	Pantograph	2	84	I							1								1	1	1		1			3			1				1	1	4	2	2		1	1	1							1	1	60																																		
50	060202	Cyl grinder	3	86	I									2	6		2	1	1	2	2	1		1	1		8		1	1				8	2								1		2			14		1	29																																
51	040501	Mill univ	1	15	C					1					2		1				1						1					2									2		2		1				2																															
52	010129	Lathe capstan	1	18	C																						2									2			2		1								1				10																														
53	020111	Lathe revolver	1	36	C																																			1			1										1	33																													
54	040406	Mill vert NC	1	38	C																														1	1			1			7		1		1		2	1							83																												
55	040101	Mill horiz Sm	3	41	C																						1							1	1		1								1			3			3			1		30																											
56	010125	Lathe centre	1	43	C																			1			1			1			1				1			8	3	15		1							1		1				7																										
57	040401	Mill vert	1	47	C			2							3		2													1	1				1	1		2				1		4		2					1					2			27																									
58	050306	Drill 4 spind	1	50	C							1	2											1									1	1		3	2									1	3	2	1					3			29																									
59	040408	Mill vert big	1	55	C													1		1			2			1		1			1	1						2				1				3		5			3					1			3		29																								
60	050201	Drill rad big	2	55	C														1		1					2		2			1			1	1	1			1										1		1		1			1			1		4	34																						
61	040407	Mill vert med	1	65	C															1	1	1											7	1	2						1							1	4					1				4		2	39																					
62	020107	Lathe capstan	1	73	C				1																3		4															1								2	6					7			3				47																				
63	050301	Drill 6 spind	1	73	C										2									2		2						2	2	3		3			1			1		3			2		3	1			2		4		1	1			1	10	24																			
64	051003	Drill capstan	1	74	C																																1	1	2					1		1			2	1	2									2		1	1		6	53																		
65	050801	Drill coor big	1	87	C								1				1			4	1		1			1			2		4			6	1		2		4							2					1	4	2			3			4		1	2	2	1			37																	
66	010707	Turn ctr	1	88	C							2		2										1		2									1	3		4	4	2	1			1			1	2	2	6		1			5			4			1	1	1	1		40																
67	010105	Lathe ctr long	2	92	C			2	1	1			1		4		3	3	2	3	1	5	2	1		4			7		2			6	2	3			5									3			2	2	2			1					2				1	3		18															
68	010127	Lathe bar/chuck	1	93	C																												5		3		4			16	7	9	2	1						1				1										1					43															
69	010123	Lathe ctr	1	109	C																1			1						1			4	1						14	18	13	21							1	3	1																	1	29														
70	051003	Drill & tap	1	117	C							1				1				2									1		2			2	2	2	2					1		1		1		1	4		3					1			1	2			3	1	2	2		3		1		75												
71	040512	Mill univ NC	2	133	C																2						1			1			2				2			6	7	18	2	1						5				4		2		3		2										16	1	2	56											
72	010704	Turning ctr	2	147	C							1	1		1					1					1		3		1			1		10	5		5	2	2	4		19		3		2		1	13	4	1	3					2			5	1	1	3			4	2	3	1			8		33										
73	051101	Tapper v. small	2	161	C							1		1	1			1		1	1				3		2							1	9	1	7			1			2		3			3	3	6	3			2	1	4			5	2		5	6	2	7		3			4		1	69									
74	051001	Drill coor NC	1	174	C																													3	1	4		2								1				1	8	1					3			4	1		9		5	1	2		1					7	9	112								
75	010119	Lahte ctr sheet	2	217	C							2		1	1	1	3	1	1	4	1	1		2	6		4		2	2		2		8		3		4			3		1		2		1	16		5	12				1	4	1		6			2	2	2	2	2	12	1			6		14	1	4	68							
76	050902	Drill coor	2	230	C													1			2			1	4		2		1	1						8	1	1				2	1		6			4	1	11	6					5						2	5	3	3		11	1			3		4	79		9	82						
77	040510	Mill univ	2	247	C																			1	1				1			1	1		1	8			17	19	30				1				13	7	1		3	2			3		2			1		1			1		9	15		2						76					
78	040504	Mill univ sm	7	272	C							2			2			1		4	2	1		2	2		5			2						8	5	9			1			4		6			4	3	12	8				4	4			12		6	5	5	9	12		8	3			17		8	34	10	17	28		7					
79	010126	Lathe ctr	7	387	C									1						1	1			2	3		2			2		2	14	1		1				49	39	30	23	1						5	2	2		4	7			4		2			1		1		11	1		19	28	1	3						11		113				
80	170502	Saw band	1	2	G							1																																																1				1																			
81	070601	Hand grinder	3	8	G							1		2						1	1										1		1				3	1																										1					1						1						4	
82	070502	Linisher	6	243	G							1		1										2			1			1	2				20	11			1				3				2	10	6				1	11		1	9	8	4	10	12	3	7	7	6	2			9	3	3	11	12	5	13		6				67	
83	170104	Saw circ.	2	526	G																		1		1																31	13	5					2	5	20					10	6	2	8	7	4	2	16		9	25	11				37	28	5	47	7	1	46	26	2	3			15	94

Note: Modules 1 to 4 inclusive and their key machines were eliminated early in the analysis.
Figures in matrix show (f) = No. of parts with ops. on each machine

50% parts
26% modules

Figure 15.25 Module summary

In choosing the groups, one starts with a scale check. In practice it has been found that groups employing 10–15 people, with perhaps one or two larger groups for complex parts, can usually be formed without having to buy any new plant. If then an existing machine shop employs 90 people, one would expect to form between six and nine groups.

Knowing the number of groups required, one finds likely groups in two main ways. First by direct examination of the modules and the module summary, and second on the basis of past experience.

In the first case, direct examination will usually find small sets of modules which use the same S and I machines, providing one or two nuclei for likely groups. Direct examination can also be used to find one or two simple groups, using only common machines to make simple parts. In the second case, experience has shown that some types of part have formed groups in many different factories. If such parts are required in reasonable quantities, it is likely that they will again form groups. Examples of this second case are: small rotational parts made on bar lathes; gears; heavy machine beds, gear cases, etc.; and simple turned milled and drilled parts. In this case it is not difficult to find one or two modules to form the nucleus of each group.

The computer is used to find the best-fit group for each of the remaining unallocated modules, and the machines are allocated to the groups in proportion to the number of parts they machine. This step finds large numbers of exceptions, most of which can be eliminated as shown in *Figure 15.26*, by the re-allocation of operations to other similar machine types which are available in the groups. There may then be one or two difficult exceptions which have to be replanned, and a final load check is advisable, to ensure that machines required in more than one group are not overloaded in one of them.

Line analysis (LA) is then used to provide the information needed for plant layout, and tooling analysis (TA) is then used to plan the group tooling and to find 'tooling families'.

15.9 Assembly groups

Assembly groups are similar to group technology component processing groups, in that both types complete all the components they make, and both are dedicated to simple material flow systems and to the minimisation of throughput times.

Component processing and assembly groups have different historical backgrounds. The group technology component processing groups were developed by engineers to improve production efficiency. Assembly groups owe their origin to behavioural scientists and were introduced originally to improve workers' morale and job satisfaction.

Research carried out by the Tavistock Institute of Human Relations in the late 1940s had shown that workers preferred to work together as a team to complete products, rather than work alone as specialists in particular processes. In the early 1950s, the Norwegian government was rebuilding their industry after the war. They decided that they wanted to introduce group working and engaged two scientists (Trist and Emery) from the Tavistock Institute and one from Norway (Thorsrud) to introduce groups into Norwegian factories. Together they formed groups in over 100 factories. They had some successes, but failed to gain the full potential advantages of the change. The experiment ran out of steam and was dropped, but it generated a wide interest in group working throughout the industrial world.

In Sweden about this time, Volvo were having difficulties in recruiting Swedes to work on their assembly lines. They brought in foreigners, mainly from Finland, to do this work. Gullenhamer, the president of Volvo, believed that there could be no long-term future for a company which had to bring in foreigners to do work which its own nationals refused. He instructed his production engineeers to study the work done in Norway, and to design an assembly factory for Volvo cars where Swedes would be happy to work. The result was the Volvo assembly plant at Kalmar, based on group assembly.

Other companies which were experiencing problems with long machine-paced assembly lines and who turned to groups at this time, included Philips in Holland, making television sets, and Olivetti making calculators in Italy. It was expected at the time that the change to groups would increase costs, but it was seen as necessary in order to eliminate the alienation of labour. In the event, the change in nearly every case reduced costs rather than increased them.

15.9.1 Planning assembly groups

There are five main ways of dividing assembly work into groups. These are illustrated in *Figure 15.27*. The first three types of group complete all the products they assemble. In the last two types—branched groups and groups in series—each group completes a major part of the product, and a number of groups work together to complete the finished product. It should be noted that even in the case of branched groups and groups in series, all assembly groups produce continuously. They do not assemble intermittently in batches for stock at any stage in assembly.

Inside the groups illustrated in *Figure 15.27* there are again five different ways in which the work can be organised. These are illustrated in *Figure 15.28*. With mono working, each worker completes products. With parallel working, several workers each complete the same type of product. With team, branched and line working, workers collaborate to complete products.

A number of groups from practice are illustrated in *Figure 15.29*. The first shows groups making television sets at one of the Philips factories at Eindhoven, Holland. Groups in parallel each made the same set. Inside the groups the work was organised for line flow. The second illustration shows groups making electronic calculators at an Olivetti factory at Ivrea, Italy. The groups were designed as branched groups, one assembling the printer unit, one the keyboard, one finishing the product, and a final group for testing. Inside the groups, some used mono working, and others used parallel and line working.

Finally, *Figure 15.29(c)* shows three of the 21 groups in the Volvo, Kalmar car assembly plant. Each group completed a logical part of the assembly, e.g. the suspension, the brakes, engine installation, electrics, upholstery, or doors, etc. The groups are organised internally for branched work, some workers making subassemblies, and others installing them on the cars. A buffer stock of up to three cars is allowed between groups, to enable groups to vary their pace of work and to take occasional breaks. To improve group team spirit, each group has its own rest room (coffee bar), showers, toilets, cloakroom and a sauna bath.

The design of assembly groups is not as complicated as the design of component processing groups. A special technique like production flow analysis is not necessary for planning. All one needs in designing assembly groups is to remember the objectives of completing products, minimum throughput time, minimum stocks and continuous flow.

15.9.2 Traditional assembly organisation

If one goes back far enough in time, factories were small and all assembly was based on mono groups. In the last 60 years,

16 Manufacturing Strategy

J Hunt

Contents

16.1 Manufacturing strategy and organisation 16/3
16.1.1 The changing environment in world manufacturing 16/3
16.1.2 The 1980s environment and strategies 16/4
16.1.3 Future strategies 16/5
16.1.4 Key objectives of a manufacturing strategy 16/5
16.1.5 Key elements of a manufacturing strategy 16/5
16.1.6 The manufacturing organisation structure 16/5

16.2 Strategies for increasing manufacturing competitiveness 16/7
16.2.1 World-class manufacturing 16/7
16.2.2 Quality-improvement programmes 16/7
16.2.3 Just-in-time processing 16/7
16.2.4 Concurrent engineering 16/8
16.2.5 Design for manufacture and assembly 16/10

16.1 Manufacturing strategy and organisation

16.1.1 The changing environment in world manufacturing

World trade has become increasingly competitive with the development of Japanese and other Asian manufacturing capabilities, the introduction of new technologies and vastly improved communications and transportation methods. In this situation, the implementation of a modern manufacturing strategy and organisation is central to the survival of a manufacturing company. Strategies and organisation structures for the 1990s and beyond will have to be fundamentally different from those of the past because of these developments.

The 1980s heralded a vast change in Western manufacturing philosophy, strategy and organisation. The continued increasing success of Japanese manufacturing in world markets plus generally fiercer international competition intruding into previously secure home markets has led Western manufacturing industry to adopt many Japanese manufacturing approaches. At the same time, improvements in UK labour relations and attitudes have smoothed the way for radical changes in manufacturing organisation and the greater involvement of all staff in the decision-making processes.

In addition, advances in technology and materials have created new industries (e.g. plastics, electronics and computers) and eliminated or severely reduced others. Existing industries that have survived the technological changes are facing an increased pace of competition. These factors, together with greater demands for product variety and higher quality, have led many leading companies to adopt new corporate and manufacturing strategies and organisation structures.

Table 16.1 shows a comparison between the general market situation for manufactured goods in the 1990s and that prior to the 1980s and the oil crisis of the 1970s. *Table 16.2* shows the

Table 16.1 The market environment

Aspect	*Pre-1980s*	*1990s*
Markets	Mainly domestic, some international	EEC per 1992, and world-wide
Source of competition	Other domestic companies, some European and North American	EEC, North America, Japan and Far East
Product range	Basic ranges of new products	More models and options needed to make sales in more competitive world markets
Product life	Varies with product, but typically 5–10 years	Varies with product, but typically 3–5 years
Consumer demand	Pre-oil-crisis boom as disposable incomes rose continuously	Prices slackening demand, plus some saturation
Industrial demand	Pre-oil-crisis; high demand on back of consumer demand	World overcapacity in many industries leads to low demand for capital goods
Quality expectation	Variable only in that demanded and achieved	Greater expectation of top quality by customer
Price demand	Booming Western economies could absorb rising costs and inflation	Stagnant economies sensitive to rising costs and inflation

Table 16.2 The manufacturing environment

Aspect	*Pre-1980s*	*1990s*
Product range	Increasing and high volumes	More model options with lower volumes
Product design	Manual drafting and early CAD	Wider use of CAD, CAE and CADCAM. Networking of design systems
Levels of volume	Steadily increasing as world demand expanded	Static, fluctuating or reducing with demand and technology changes
Volume production methods	Dedicated flow lines and machines	Flexible manufacturing systems
Batch production methods	Stand-alone machines	Flexible cells and JIT processing
Production technology	Single-purpose machines. Electromechanical controls	Multipurpose machines, CNC, computer controls
Manpower levels	Labour intensive	Reduction of both direct and overhead labour
Manpower skills needed	Single skill needed for one job only	Several skills needed for multiple tasks
Specialist skills	Organised in central service departments	Devolvement to the user departments
Quality assurance	Variable—only military and aerospace subject to quality-assurance systems	Universal demand by customers for implementation of quality assurance systems to ISO 9000; quality improvement programmes
Quality control	Inspection by roving inspectors and a central quality-control department	SPC and operator inspection/responsibility
Inventory policy	Large batch, make for stock, central stores	No stock. Make when needed (JIT)
Lead times	Variable with industry, but 12–16 weeks typical	Variable with industry, but 2–4 weeks demanded
Data processing, information technology	Centralised in data-processing department	Distributed to users as local processing power increases
Organisation structure	Hierarchical management with many levels	Team working with fewer management levels

CAD, computer-aided design; CAE, computer-aided engineering; CADCAM, computer-aided design and manufacture; JIT, just-in-time; CNC, computerised numerical control; SPC, statistical process control.

same comparison for the manufacturing environment. The comparisons are in the form of generalised statements that will apply to most industries and companies, with perhaps (but not necessarily) the exception of those in a strong monopoly or protected position.

The current situation is summarised below.

(1) Product competition is forcing the faster introduction of new product designs which will have a shorter life cycle than their predecessors.
(2) Price competition is forcing companies to be ever more cost efficient.
(3) Quality competition is demanding better working practices, methods and controls.
(4) Markets for manufactured goods have moved away from relatively insular domestic products and markets with prospects for continuous growth, to a world market of fluctuating demand, shorter product lives and fierce international competition invading domestic markets. The effect of this on the UK is illustrated in *Figure 16.1*. This figure shows the decline in the balance of UK exports over imports of manufactured goods from 1978 to 1989. In 1989 the UK had gone to a deficit of nearly £17 billion. The negative balance is now three times the positive balance at the start of the 1980s.
(5) The opening up of Eastern Europe and the continued development of the EEC will consolidate the total European market and effectively eliminate the UK as an isolated and separately supportable market for most products.

16.1.2 The 1980s environment and strategies

The overall effect of the factors described in *Table 16.1* led in many cases to sales being limited by productive capacity. The resulting supply-driven (i.e. production-driven) market gave manufacturing management the justification for employing more labour as the short-term solution for increasing production capacity. This in turn created extensive hierarchical management structures to control a large labour force which was organised to do single repetitive tasks.

As technological advances came into place (plastics, microelectronics, computers, and computerised numerical control (CNC) machining), additional specialist staff were required to apply the technologies. The specialists were usually organised into central service departments that could provide a service to all the 'user' departments.

The combination of a multilevel hierarchal management and central support services took nearly all responsibility and motivation away from local supervision and workforce. Despite this, the system seemed to work well because the continuous general high level of demand provided enough revenue to pay for extensive managerial and specialist resources. However, companies started to see severe problems when:

(1) demand dropped through (a) the general market situation, and (b) the oil crisis of the 1970s;
(2) the competitiveness and new technologies of Japanese products took markets away from European and North American companies; and
(3) the consumer (both private and industrial) started to demand more variety and better quality in addition to price competitiveness, i.e. more value for money.

A partial solution to these factors was seen in the 1970s with the introduction of group technology. This is primarily an organisational change, creating a multitasked group or cell to carry out all operations on a component or assembly instead of feeding them through a series of separate departments of like machines (lathes, mills, drills, etc). This served to reduce lead times dramatically but, of equal if not greater importance, also gave some responsibility back to workers in their role as members of the cell team.

Group technology, although successful in itself, was still operated within the old structures of the typical manufacturing organisation and, as a consequence, failed to reach its full potential in many cases. However, its basic principles of cell

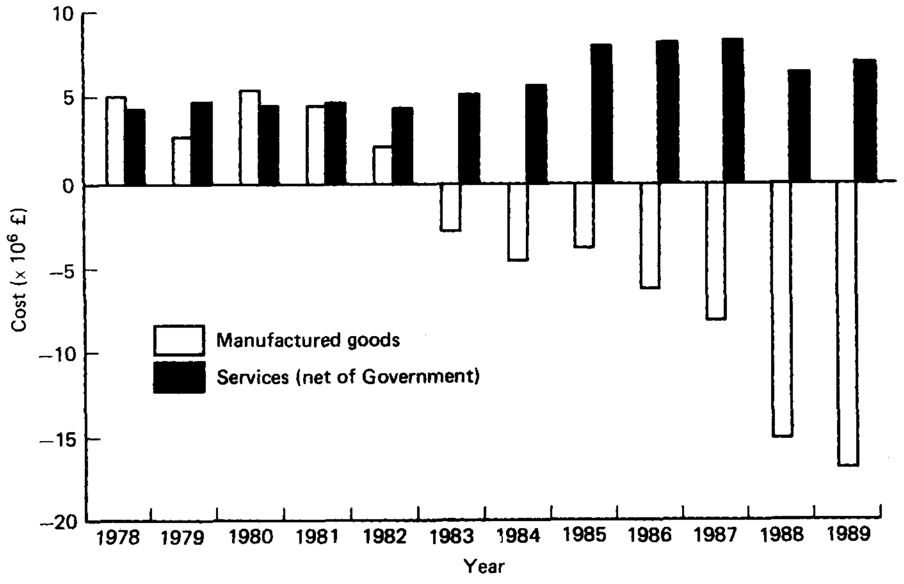


Figure 16.1 Balance of trade in manufactured goods and services. (Courtesy of CSO Pink Book)

manufacture, team working and wider responsibility for the individual worker are fundamental components of the new manufacturing strategies that are now emerging.

16.1.3 Future strategies

The general trend now seen as we move towards 2000 is being expressed in a number of strategies and techniques (see *Table 16.3*). It is arguable whether some of these are strategies in the sense of being high-level policies or operational techniques applicable at departmental level. Such definition is of minor interest, the more important point is that they are a group of activities that can be applied across a manufacturing organisation, and bring about a major improvement in competitiveness.

Many of the strategies listed in *Table 16.3* are already being implemented in companies—some at departmental and local levels and others such as world-class manufacturing (WCM) or computer-integrated manufacture (CIM) at entity level.

16.1.4 Key objectives of a manufacturing strategy

When considering a manufacturing strategy, management should keep in mind several key objectives that are at the centre of most modern strategies. These are:

(1) *Integration*—structure the total manufacturing system to work as one unit and to one well-understood strategy.

Table 16.3 Modern manufacturing strategies and techniques. There are a very large number of strategies, philosophies and techniques being used as methods of increasing competitiveness in cost, quality and delivery to the customer. Some of the more widely applied are given in the table

	Acronym
Enterprise organisation	
World-class manufacturing	WCM
Computer-integrated manufacture	CIM
Electronic data interchange	EDI
Networking—wide and local areas	WAN or LAN
Engineering	
Computer-aided engineering	CAE
Computer-aided design and manufacture	CADCAM
Concurrent or simultaneous engineering	CE/SE
Design for manufacture and assembly	DFMA
Engineering data management	EDM
Production	
Computer-aided production management	CAPM
Materials-requirements planning	MRP
Manufacturing-resource planning	MRPII
Optimised production technology	OPT
Just-in-time	JIT
Flexible manufacturing systems	FMS
Quality	
Company-wide quality improvement programs	QIP/CQP/CQI
Total quality management	TQM
Statistical process control	SPC
International quality system standards	ISO 9000/BS 5750

(2) *Flexibility*—create a situation where the minimum of hardware and human resources are capable of handling multiple products and operational demands.
(3) *Involvement*—involve staff more fully in consultation and decision-making in aspects that effect their work. This is a predominant feature of quality-improvement programmes, but should be generally applied as a company philosophy and manufacturing strategy.
(4) *Quality*—endow the complete organisation with a 'quality culture' that involves every individual in a continual quest for the improved quality of their performance and output.
(5) *Inventory*—create a situation where varying demand can be met from the lowest possible inventory.

It will be noted that measured cost and efficiency do not appear in these objectives. The philosophy behind this is that if the organisation can achieve a culture that meets the above objectives then lower costs and greater efficiency should come.

16.1.5 Key elements of a manufacturing strategy

When setting a manufacturing strategy the following elements need to be considered. Some are common to most manufacturing organisations, but the emphasis will vary from company to company.

The elements to consider include the following.

(1) Basic nature of operation—assembly only, full manufacturing of piece parts, subcontract content.
(2) Production volumes and mix.
(3) Inventory—inventory carrying policy and value targets; materials sources and costs.
(4) Manufacturing resources—what is needed after consideration of the above elements?
(5) Labour availability—requirements and availability of staff and their skills.
(6) Location—alternative sites for manufacture in relation to overall profitability, markets, available resources and investment needed.
(7) Existing product range—profitability, market life remaining, cut-off point and continued spares/service support.
(8) New products—development costs, potential profitability, lead time to market, product life cycle, distribution, service and spares.

16.1.6 The manufacturing organisation structure

16.1.6.1 Past structures

As noted before, most companies used to operate in a hierarchal way through many layers of management. *Figure 16.2(a)* shows a typical functional structure with the flow of product at 90° to the flow of responsibility up and down the hierarchy. Decisions, data and responsibility for all activities go up and down the organisational structure. This is not to say that interdepartmental communications do not take place at every level, but the need for say a manufacturing engineer, production supervisor and quality engineer to refer up the line and wait for separate responses in addition to trying to solve the problems at their level, does cause uncertainty and, possibly, a delayed solution.

The product flow is shown going across the page as materials flow and value is added between the goods-in and despatch doors. Systems are put in place to allow the product progress to proceed without constant reference up the management line, but there is not sufficient focus of individuals to

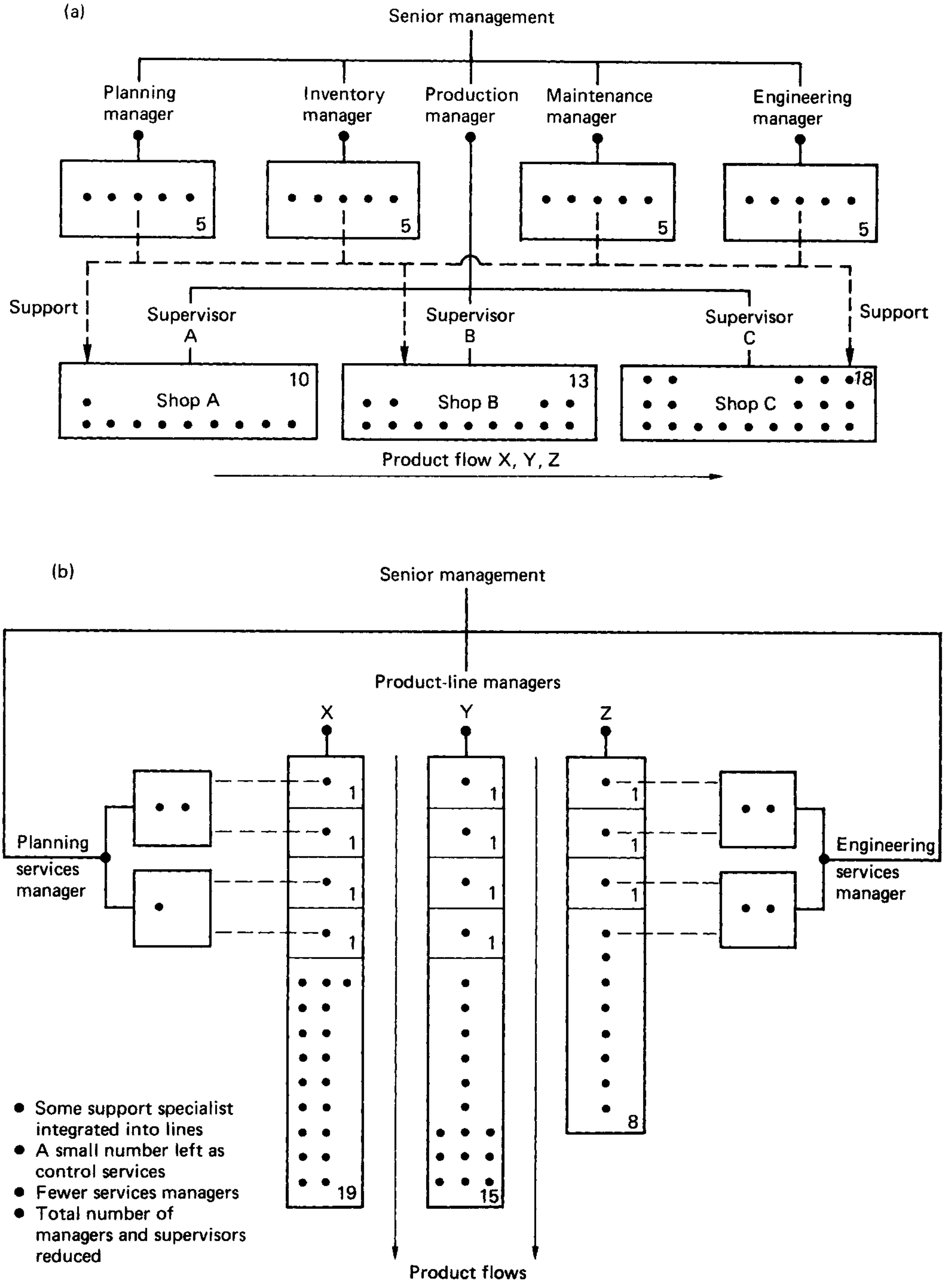


Figure 16.2 Manufacturing organisation structures. (a) Traditional organisation. (b) Product-line structure

products, except in dedicated and rather inflexible production lines. In any case the 'production line' is not the only contributor to the product flow. The focus of the centralised support and service functions is just as important and a lack of product focus in these areas is often a major block to the organisation's total efficiency door-to-door.

16.1.6.2 Future structures

Manufacturing organisation structures will be market and product oriented with the skills of the production and support staff aligned more to product groups than process groups. They will also be 'flatter' with fewer levels of management, as responsibilities are devolved back to the operating staff. *Figure 16.2(b)* shows such a structure which encourages:

(1) better communications between the individuals contributing to the design and production of the group product—giving a better and faster response to events and problems;
(2) a greater sense of 'ownership' of the product by the group members and, therefore, greater individual motivation;
(3) a greater input by individuals of possible solutions to problems; and
(4) multitasking and multiskilling becomes more acceptable to individuals in the interests of the group and their own development.

These factors will result in more highly motivated working and better quality, which in turn will create shorter lead times and lower costs.

16.2 Strategies for increasing manufacturing competitiveness

16.2.1 World-class manufacturing

World-class manufacturing (WCM) is the name given to the attainment of a level of customer service and manufacturing efficiency that ranks with the best in the world. With the increase in international competition, leading companies have realised that they will only survive if they are among the best in a world-wide market. They have also realised that in order to achieve such standing there must be a company-wide strategy and commitment to that level of attainment.

The adoption of a WCM strategy by top management leads to a commitment to vigorously pursue every avenue of improvement in the company. In particular, this is expressed as improvements in the *quality* of each person's contribution and all operational activities.

The *key objectives* of a WCM programme must be:

(1) 100% customer service/satisfaction,
(2) 100% quality in the product,
(3) competitive pricing,
(4) shortest delivery and manufacturing lead times possible, and
(5) lowest costs consistent with the above objectives.

Item (1), customer satisfaction, will come from the achievement of other objectives.

The *key philosophies* to be adopted in a WCM company relate to people and must be:

(1) every member of the company *must* adopt an attitude of personal responsibility for the highest possible quality of his/her work;
(2) the company and its departments are *teams* not warring groups or individuals;
(3) every member and working department of the company must think in terms of being a supplier to the next person or department in the chain, who must be regarded as customers; and
(4) the internal customers have as much right to the highest level of service possible as do the ultimate customers of the company.

The *operating techniques and methods* will vary from company to company, but some of the modern practices one would expect to see in use include:

(1) company-wide quality-improvement programmes;
(2) just-in-time (JIT) working;
(3) concurrent engineering;
(4) design for manufacture and assembly (DFMA); and
(5) computer-integrated manufacture (CIM)—
    (a) computer-aided production management (CAPM), and
    (b) computer-aided engineering (CAE and CAD-CAM).

For more details on item (5) the reader should consult Chapter 14 in this volume.

16.2.2 Quality improvement programmes

Quality-improvement programmes (QIPs) go under various titles. Some are proprietary products sold by consultants, whilst others are internal company schemes. They are vehicles for the application of a company's total commitment to attaining the highest level of quality throughout the organisation. QIPs are not a solution in themselves.

QIPs are designed to achieve the following.

(1) Increase each person's awareness of the need for high quality in their work.
(2) Apply QIP to all departments and activities—it is not a traditional quality inspection technique restricted to the shop floor only.
(3) Install a system that has formal procedures of working to follow, but which at the same time encourages people to use it in a personal and relaxed manner.
(4) Allow groups of people to examine their working situation and discuss and plan improvements both very minor and major.
(5) Provide a means of self-measurement of results plus local visual presentation of the plans and results.
(6) Train all people in their participation in the programme as individuals and as members of their working group. Shop-floor and office staff are trained in simple statistical and presentation techniques so that they have a greater interest through doing the analysis themselves.

16.2.3 Just-in-time (JIT) processing

Just-in-time (JIT) processing has been applied extensively in the last few years. It has proven to be an effective way of reducing lead times and inventory by working to the principles of a 'pull system', in which parts are made only as they are required by the customer process or department. In many companies, JIT has produced (at low capital cost) some of the inventory and cost reductions that high technology planning and manufacturing projects have failed to deliver. Reductions of over 50% in inventory and lead times have been achieved in many implementations.

A successful JIT implementation can be achieved by addressing the subjects of JIT philosophy, organisation, product manufacturing structures, batch sizes, and lead times.

16.2.3.1 JIT philosophy

As with most other activities, success can only be achieved by having the full commitment of the people involved. JIT is a physically visible system that is driven at shop-floor level rather being a top–down paper planning technique. In a well designed and understood JIT environment the shop-floor supervisors and staff become the owners as well as operators of the physical system. This provides a much better working atmosphere and interest because they can see the results of their own and others' work and, most importantly, make a positive contribution to the resolution of problems.

It is vital, therefore, that senior management engender the correct spirit of team work and involvement on the part of the shop-floor staff and their support services. JIT work teams should hold regular meetings (weekly if not daily) at their workplace to sort out current problems and for further education and the distribution of management information.

16.2.3.2 Organisation

The fundamental element in JIT systems is the *kanban* which is the repository of a JIT batch and may take the form of a tote bin(s), stillage(s) or other container(s) or may just be an area designated on the shop floor. The number of parts in the kanban is preset (but variable in response to situations), and becomes the process and transfer batch quantity. The number of kanbans will depend on the number of separate operations in the revised 'flatter' manufacturing structure.

JIT teams should be kept relatively small (say, less than 10 people) in order to maintain a close team spirit and sense of identity. The style of the team should be democratic with a supervisor or team leader advising, informing and guiding rather than ordering. Many JIT teams run with only a nominal hierarchical supervisor—in any well-mixed group natural leaders will emerge in different situations.

Communications should be kept simple, with paperwork at a minimum. Whiteboards, blackboards or flip charts should be used to display product and performance information in large characters. They should be sited in a prominent position within the JIT area and updated on a very frequent basis. Staff should take turns in this role to increase their involvement.

16.2.3.3 Product structures and bills of materials

A major technique for improving throughput and reducing lead times is to reduce the number of subassemblies and separate operations. This reduces the number of batches and their transfer operations, decreases work in progress (WIP) stocks and increases staff interest and their work. When introducing JIT the product structure and bill of materials (BOM) progress through manufacturing should be examined and modified to make them 'flatter', by removing unnecessary intermediate stages.

16.2.3.4 Batch sizes

Batch sizes are reduced to those that can be progressed in an appropriately small quantity. A batch of 1 may be ideal from the standpoint of having the lowest inventory value, particularly for high-value, long-process-time parts or assemblies. It is unlikely to be ideal for low-value, short-cycle processes because it creates too many demand pulls on the system, thereby increasing the time and cost of progressing relative to the actual processing time.

Nevertheless, batches should be as small as is practicable. There are no generally applicable rules for this—each product and process must be carefully examined to determine the best balance of batch quantity and process time needed in relation to the whole production process of the product and, in particular, to the preceding and succeeding operations. The key objective is to balance the *flow* of the overall product in small efficient lots. The balancing of the start-out quantities of all the WIP is to be avoided in order to reduce the excessive inventory costs caused by varying rates of WIP consumption.

16.2.3.5 Lead times

The overall lead times of most products and assemblies are determined by the longest lead time within the constituent parts. Typically, one part with a lead time of, say, 10 weeks constrains the assembly of the rest of the parts which have lead times of 1 or 2 weeks. Long lead times can often be reduced by modifying the supply chain, BOM, process order and batch size of the parts and/or of the parent assembly.

The response time of the 'on-demand' pull system greatly reduces the lead time of a part. In other words, the direct action of the supplying department to the pull-system demand can avoid the delays often found when supposedly premanufactured parts are not in stock—the task of then getting these made quickly only further upsets a factory-wide scheduling system and capacity.

16.2.4 Concurrent engineering

Concurrent engineering (also called simultaneous or parallel engineering) is the technique of simultaneously considering the design, quality, manufacturing engineering, production and sourcing implications of a new product design *during the design process* and not as a series of afterthoughts. This methodology has long been the desire of manufacturing personnel, but has often been less than welcome, and therefore unworkable, with design staff. The method has only become fashionable with the late realisation by non-manufacturing managers that it can significantly reduce the lead time to bring a new product from conception to the market place. Short lead times to market are becoming increasingly important for manufactured goods as fierce world-wide competition is demanding new or improved products at an ever-increasing rate. It is vital that this demand is met if a company is to maintain, let alone increase, its market share and profitability.

Concurrent engineering facilities are now being offered by computer systems as facilities on CADCAM and CAE software where the sharing of databases allows simultaneous or parallel working rather than serial working (where manufacturing engineering, quality engineering and production planning work started only after the substantial completion of the design work).

However, the main factor in concurrent engineering is multidisciplinary team working. The early participation of manufacturing personnel in the actual design process can make significant savings in development and production costs and lead time. Lead time is reduced by avoiding time spent on unmanufacturable or too costly designs and also in time savings through the overlapping and parallel working of the departments concerned.

A concurrent engineering programme is shown in *Figure 16.3*.

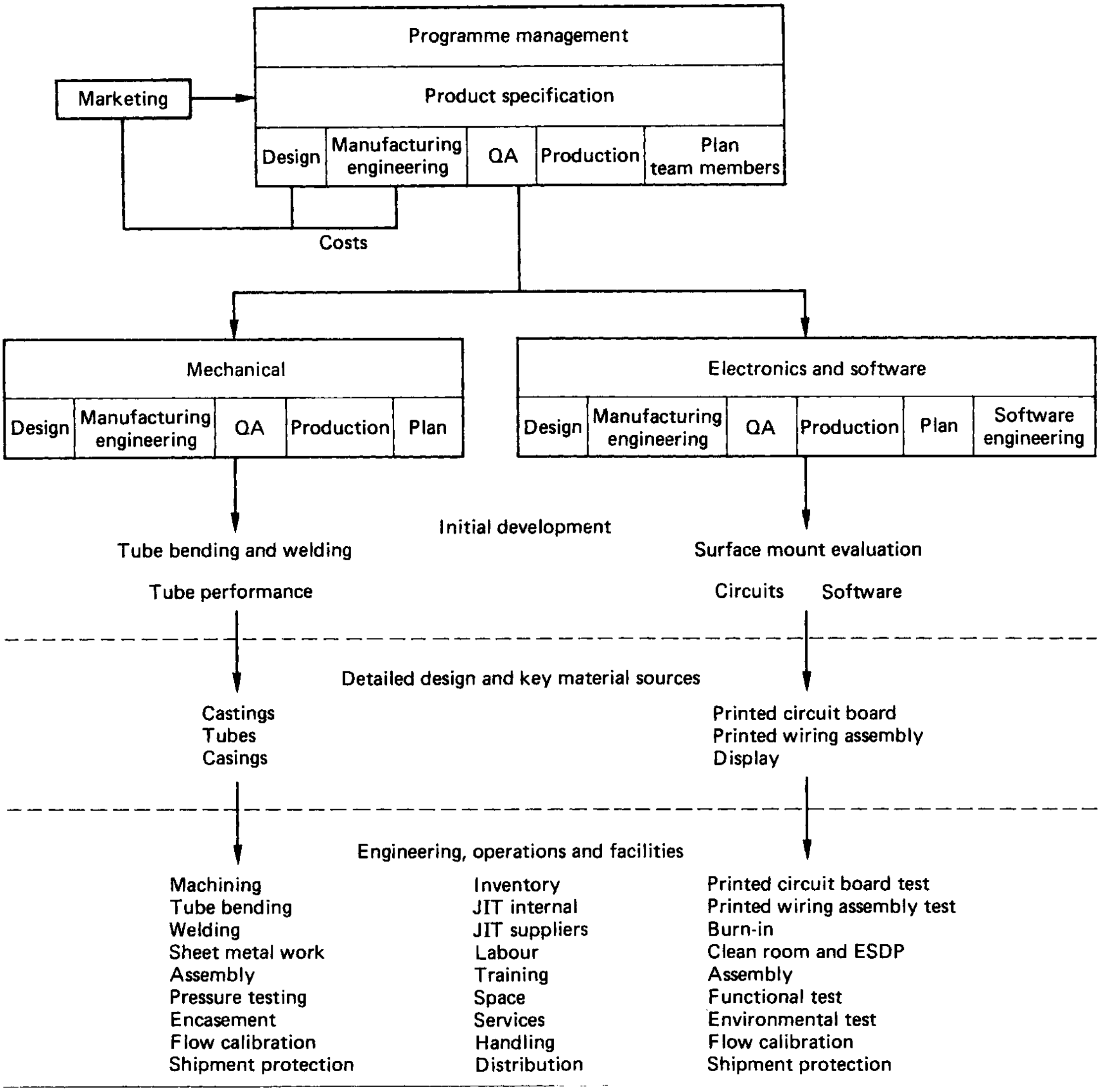


Figure 16.3 Concurrent engineering program for a combined mechanical and electronic product. JIT, just-in-time; QA, quality assurance; ESDP, electrostatic discharge protection. (Courtesy of Custom Networks Ltd)

16.2.4.1 A concurrent engineering project

Within a factory environment there are usually six functions that play a major role in the manufacture of products: production departments; manufacturing engineering; quality control; production control; inventory control; and purchasing. For successful project completion, all of these must be involved as early as is advisable.

Phase 1—Initial development and design As an absolute minimum for concurrency, an engineering team should be established comprising development engineers, product designers, manufacturing engineers and quality engineers. Being an engineering team, this group should be able to deal with all the technical considerations of the emerging design. In addition the manufacturing and quality engineers should also be able to represent the general interests of the production floor because of their close daily working with it.

Phase 2—First design review In addition to the engineers, this should include staff from production, production control and inventory control in order to gain their first formal inputs to the preliminary design.

Phase 3—Full design and preproduction planning Two teams should now be set up. Whilst the engineering team(s) continue with the more detailed design of the product, the processes and their quality requirements, a manufacturing team will start preliminary planning of sourcing and production. It is recommended that a key designer, manufacturing engineer and quality engineer should sit on the manufacturing teams, and vice versa. This stops the project and its progress relying

on one very large team meeting, but maintains the communications between the parallel engineering and manufacturing teams.

Phase 4—Final design review This may be preceded by intermediate reviews at appropriate programme or project further-authorisation points. This official authorisation step is essential if, in the rush to market, the company is to avoid using large amounts of resources in introducing a product that does not meet its performance, cost or market objectives. The review should be attended by key personnel from all the departments involved and by senior management. If the design is accepted this initiates full tooling for production and production materials sourcing, building on the preliminary work already undertaken. Because of the long lead times involved, some tooling, capital equipment and sourcing commitments may have to be made prior to full design acceptance, but this is a matter for management judgement.

16.2.5 Design for manufacture and assembly

Design for manufacture and assembly (DFMA) is a technique for analysing the cost efficiency of a proposed design during the design stage. It is a technique that has recently attracted attention in manufacturing and design circles with the realisation that unnecessary cost is best eliminated in the design work, not once a finalised design has reached the manufacturing domain. Commonly, about 90% of product cost is committed at the design stage (as logically it must be).

Manufacturing engineers have long wished for a greater consideration of manufacturability in the design but, as with concurrent engineering, it has taken fiercer world-wide competition to bring the message home to senior management and product designers. Many companies now operate a system of reviewing designs for manufacturability and cost efficiency. This involves a measure of concurrent engineering practice, but may well be performed under varying titles and engineering organisation structures.

Interest has also increased with the development of computer program versions of the basic design analysis methods. These methods address design and manufacturing factors and use a points or ranking system to evaluate them. The factors considered include:

(1) product and parts—output volumes, mix, and families;
(2) what production processes exist for part manufacture—cost of parts per process;
(3) part symmetry and other geometric features—ease of grasp and alignment; and
(4) type of assembly joining—snap fit, screw, bolt, adhesives.

The systems have libraries of data on which to draw as the user is taken sequentially through a question/answer or statement menu. The results can give guidance not only on the most efficient product cost, but also on the process and capital costs implied.

17 Control of Quality

H L Cather

Contents

17.1 The concept of quality 17/3
- 17.1.1 Defining quality 17/3
- 17.1.2 Quality—a competitive weapon 17/3
- 17.1.3 Quality costs 17/5
- 17.1.4 Quality and the law 17/6
- 17.1.5 Organisation and the duties of a quality-assurance section 17/7

17.2 Quality through integrated design 17/7
- 17.2.1 The role of design with regard to quality 17/7
- 17.2.2 Specifications 17/7
- 17.2.3 Modifications to design 17/7
- 17.2.4 Process selection 17/8
- 17.2.5 Design techniques to aid quality 17/8

17.3 Standards 17/10
- 17.3.1 British Standards Institution 17/10
- 17.3.2 Quality systems (BS 5750/ISO 9000) 17/12

17.4 Material and process control 17/13
- 17.4.1 Incoming material 17/13
- 17.4.2 In-process control 17/15

17.5 Process capability 17/17
- 17.5.1 Process variability 17/17
- 17.5.2 Problem identification and highlighting techniques 17/18

17.6 Product acceptance sampling schemes 17/20
- 17.6.1 Product inspection frequency 17/20
- 17.6.2 Acceptance sampling schemes 17/21
- 17.6.3 Factors affecting the choice of sampling plan 17/22
- 17.6.4 Use of BS 6001 sampling procedures and tables 17/22
- 17.6.5 The operating-characteristic curve 17/23
- 17.6.6 Calculating a non-standard quality plan 17/25

17.7 Statistical process control 17/26
- 17.7.1 The normal distribution 17/26
- 17.7.2 Variable control charts 17/26
- 17.7.3 Attribute control charts 17/30

17.8 Measurement of form and surface 17/32
- 17.8.1 National measurement system 17/33
- 17.8.2 Characteristics of measuring instruments 17/33
- 17.8.3 Linear measurement 17/34
- 17.8.4 Angular measurement 17/35
- 17.8.5 Limit gauges 17/36
- 17.8.6 Surface texture 17/38
- 17.8.7 Circular (roundness) measurement 17/39
- 17.8.8 Air gauges 17/41
- 17.8.9 Electronic instruments 17/41
- 17.8.10 Inspection probes 17/41
- 17.8.11 Coordinate measuring machines 17/42
- 17.8.12 Optical systems 17/43
- 17.8.13 Measurement of thread pitch 17/43

17.9 Non-destructive testing 17/46
- 17.9.1 Radiography 17/46
- 17.9.2 Ultrasonics 17/50
- 17.9.3 Magnetic particle testing 17/52
- 17.9.4 Penetrant testing 17/54
- 17.9.5 Eddy-current testing 17/55
- 17.9.6 Electropotential crack depth detection 17/56

17.1 The concept of quality

17.1.1 Defining quality

The word 'quality' is used in several different contexts which can lead to the meaning becoming slightly blurred, the last thing that is desired in quality. The mainstay of all matters relating to quality are the international and national standards, so where better to look for a definition of quality than in a British Standard?

BS 4778: 1971 is a British Standard devoted to a glossary of terms used in quality assurance. Within this standard, the following definitions are recommended to be used to differentiate between the three most common usages of the word 'quality':

Quality The totality of features and characteristics of a product, or service, which bears upon its ability to satisfy a given need, i.e. a fitness of purpose, or value for money, sense.

Grade When applied to a material or product, the quality grade is an indication of the degree of refinement; when applied to a service, the diversity of functions or facilities offered, i.e. this is a comparative or degree-of-excellence sense.

Quality level This is a general indication of the extent of departure from the ideal, usually a numerical value indicating either the degree of conformity or non-conformity of a product to its specification, i.e. a qualitative sense.

In order to control or improve quality, it is first necessary to identify, measure and evaluate it. The above definition of 'quality' calls for the identification of the characteristics and features bearing upon the fitness of purpose of a product or service. Furthermore, the ability to satisfy a given need must include economics as well as design, availability, maintainability, and all other characteristics that the need for the product or service involves.

This demonstrates that the quality is determined by all the parts of a business acting together. Quality can then be said to be the sum of

(1) recognising the customers' needs;
(2) designing products, or services, to meet them;
(3) producing reliable products to promised delivery dates; and
(4) maintaining effective after-sales service and feedback.

These elements add up to fitness for purpose, a perceived value for money and hence satisfied customers.

Figure 17.1 shows some of the determinants and measures of the quality of a product that bear upon its ability to satisfy a given need. Note that even in this figure, which is an extract from BS 4778: 1971 the term 'reliability' is used in two different senses—firstly in a general sense and then to denote particular characteristics which can be stated in quantitive terms.

It is important that quality is not only applicable to tangible artefacts and services, but also to verbal and written information.

17.1.2 Quality—a competitive weapon

In every market, an organisation has to operate against keen competition in order to gain its share of the available orders. It is, therefore, important that each organisation clearly identifies the criteria which will enable it to enter a market and then win orders against its competitors. The common order-winning criteria are:

(1) price,
(2) design,
(3) quality,
(4) product range,
(5) delivery speed,
(6) technical liaison, and
(7) delivery reliability.

Price by itself is seldom sufficient, therefore non-price factors such as quality are key determinants of an organisation's current order book and its survival into the future. It is very important that an organisation becomes able to assure prospective customers that their needs will be met by products

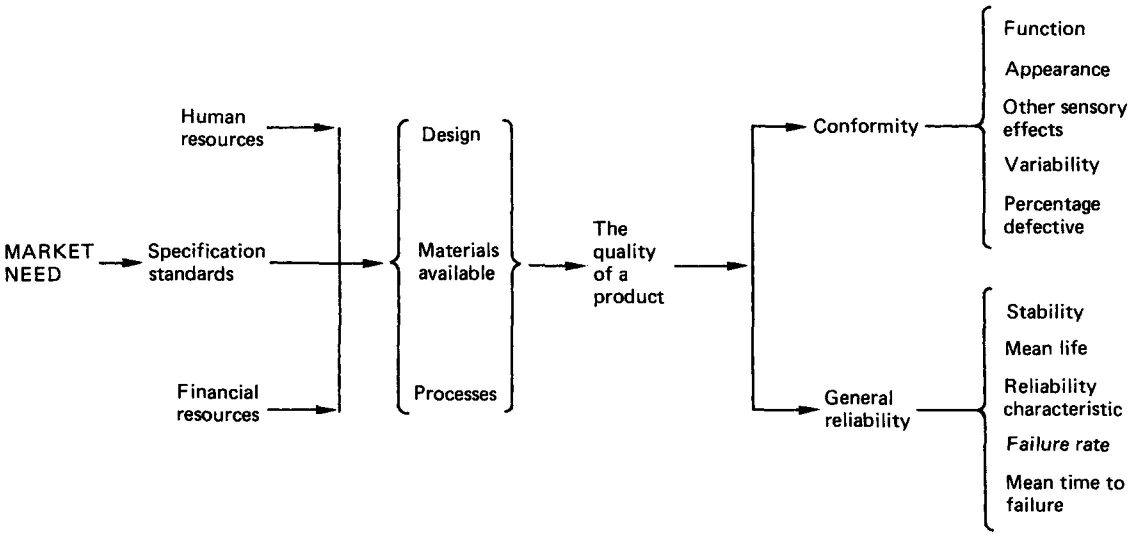


Figure 17.1 Some of the determinants and measures of the quality of a product

which are fit for their purpose. This assurance should not just be in the narrow sense of provision of proof of attainment of a particular quality level, but should be readily perceived by the customer.

Quality is of prime importance to a customer and few suppliers will not claim to be striving for good quality. Most, in fact, spend considerable time and resources on activities such as testing and inspection to detect and correct defective products. In addition, they have a large section which can speedily deal with any problems found during operational service. However, this effort, by itself, cannot ensure that quality is guaranteed. It is a form of quality control—a technique which sustains product quality to the specified requirements. Quality control is a subset of quality assurance, the latter being the key to efficient attainment of quality.

Quality assurance involves the whole of an organisation and covers all the activities concerned with the attainment of the product's quality. BS 4891: 1971 *A Guide to Quality Assurance*, gives the basic elements involved in quality assurance which are set out in *Table 17.1* and briefly described below.

Management objectives These should be closely linked to each market where an identifiable need exists. This requires an examination to determine for each product range the market entry and selling factors relating to quality, price, reliability, delivery, etc. Top management must give a clear lead to the importance attached to quality assurance by the organisation. Key personnel should be identified to resolve all quality-related matters.

Table 17.1 The basic elements involved in quality assurance

	Market demand
Design specification control	Management
Specification	responsibility
Design engineering	for
Value analysis	QUALITY
Reliability engineering	ASSURANCE
Development engineering	
Documentation	usually
Design review	involves
	defining
Purchasing control	
Purchasing	MANAGEMENT
Vendor appraisal	OBJECTIVES
Vendor surveillance	establishing a
Receiving inspection	
	PROGRAMME &
Manufacturing control	PLANNING
Work study	operating a
Process planning	
Resource allocation	QUALITY
Tooling and gauges	CONTROL
Material processing	SYSTEM
Production and assembly	taking
Inspection and testing	advantage of
Packaging	inherent
Storage	knowledge
Shipping	and/or
	research and
Marketing and servicing	development,
Maintenance	subcontracting
Product support	subject to
After-sales service	
Market reaction	REVIEW &
Complaints	EVALUATION

Programme This sets out the steps to be followed in applying quality assurance throughout an organisation from the initial product conception to after-sales service.

Planning This seeks out any special requirements which may arise and lays down the procedures for dealing with any that do so.

Quality control This establishes the standards that will be met, the method of measuring comformity with these standards and the action to be taken when the standards are not met.

Design control Quality must be designed into the product to meet the customers' requirements through a design that allows practicable and economic manufacture and servicing. The design must encompass:

(1) minimum functional/operational requirements,
(2) maximum environmental expectations,
(3) technical and manufacturing limitations,
(4) safety requirements, and
(5) cost limitations.

Purchasing control All organisations are responsible for the outgoing quality of their products and they must, therefore, ensure that all bought-in material contributes towards meeting the customers' needs.

Manufacturing control Manufacturing is where the specification is finally translated into a finished article. All processing should therefore be carried out in controlled conditions to detailed instructions.

Material control This establishes a system to identify, segregate and protect all material from initial entry into the plant until reaching the end-user, including during transportation.

Marketing This is a direct link to the customers which should continually seek out and feed back market reactions. In particular marketing should:

(1) identify market requirements,
(2) consult manufacturing before accepting non-standard work or quality level, and
(3) ensure that all costs are considered in pricing products.

Servicing In addition to providing prompt efficient servicing and attention to complaints, servicing is responsible for the provision of technical documentation, user training and spares provision.

Effective, continuous market research is the starting point for achieving quality in the eye of the customer. Unless the customers' needs are known, and understood, a designer cannot meet them.

The designer must also ensure that the manufacturing processes are considered to ensure that the specification can be reached and held. The designer should also work closely with material and component suppliers to ensure that these also meet the specification requirement.

The entire manufacturing function must be committed to meeting the design specification each and every time. Quality must be built in, it cannot be inserted later by inspection.

Many customers cannot protect themselves by inspecting fully the finished product. It such cases, the supplier must

provide some objective evidence that the specification has been achieved. This can be by test reports, or by guaranteeing adherence to an established national, or international, standard.

Marketing reinforces the customers' confidence in the product by back-up in technical liasion, servicing and spares availability.

17.1.3 Quality costs

The total of the quality-related costs within a company is very large. It has been estimated that in the UK some 10% of the Gross National Product can be classified as being quality related. Within any organisation they will range from 5% to 25% of the sales turnover value.

These costs are mainly due to failures and appraisal which add nothing to the real value of a product. They are reflected in additional costs and hence lower profits to every organisation which has them.

Many of these costs are preventable to some degree but, because they are not correctly identified, they often do not attract sufficient management input into this prevention.

There are considerable problems in clearly identifying all the costs which are quality related. BS 6143: 1992 *Guide to the Determination and Use of Quality Related Costs* does, however, identify the following four simple classifications.

External failure costs These costs arise from the failure of the product outside the organisation. They are identified by items such as complaints, warranty work, returns, etc. In many cases they result in the loss of future customers, either through their direct experience or from passed on experiences of others. It is estimated to cost 10 times as much to gain a new customer as to hold onto an existing one, and hence this can be fatal.

Internal failure costs These are costs arising inside the organisation by failure to meet the specified quality. They are identified by costs such as scrap, lost time, reworking, etc.

Appraisal costs These are the basic costs of assessing the quality being attained, e.g. inspection, testing, equipment, etc. Some of these could in fact be reclassified as internal failure costs as their main function is to screen out defective products.

Prevention costs These are the costs involved in any action taken to identify and eliminate the causes of quality failures. They include product design modifications, equipment maintenance and modification, vendor assessment, training, etc.

These costs are often not identified separately as quality costs by normal accounting procedures, but are submerged with other data as direct and indirect costs. Whereas the collection and reporting of these costs should not be a separate exercise from the remainder of the management accountancy information, they do require accurate separation in order to highlight them.

The quality costs can be established but are sometimes obscured. They must be separated out from the others. It is important to ensure that they carry an equitable share of overheads such as rates, heating, canteen, etc.

Figure 17.2 shows how these quality costs interrelate to one another. As more resources are committed to the prevention area, the occurrences of failure should reduce and hence their related cost should reduce. Note, however, that the reductions are often of a diminishing-returns nature. An economic balance between the prevention and failure costs can be breached by excess preventive costs as readily as by failure costs.

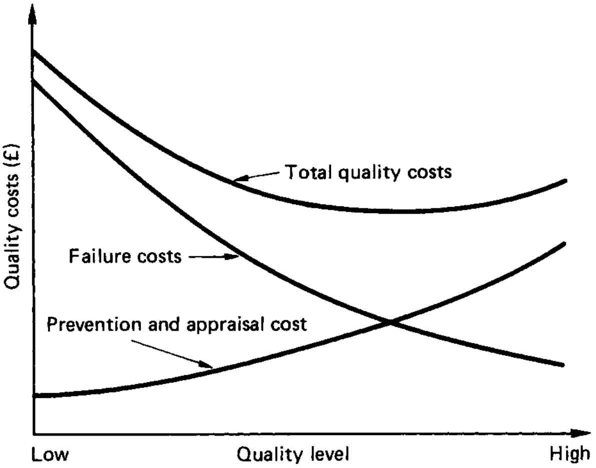


Figure 17.2 Build up of quality costs

The basis to be used when justifying expenditure on preventive action must be the same as for any other revenue expenditure or capital investment. The cost of carrying out the preventive action should result in increased profit, be it through increased sales, a higher price differential or reduced costs. Quality prevention is only worthwhile when it pays for itself.

The other main reason for identifying quality costs is control. Ideally, there should be clear standards for comparing the quality costs against, but even without standards budget control can be carried out. Variations and high costs can be readily identified and tackled.

Measuring quality costs firstly gives an indication of how large a portion of the organisation's costs arise here. They can be used as an indicator of continued performance, or improvement. They focus attention and hence aid decisions on where resources could be directed to give the best chance of good returns. More importantly, measuring is a prerequisite for control.

Reports and ratios should be set up which are relevant to the organisation's objectives and sensitive to any change in cost or business trends. The following are commonly used:

(1) Labour base: $\dfrac{\text{Internal failure costs}}{\text{Direct labour cost}}$

(2) Cost base: $\dfrac{\text{Total failure costs}}{\text{Manufacturing costs}}$

(3) Sales base: $\dfrac{\text{External failure costs}}{\text{Sales revenue}}$

(4) Product base: $\dfrac{\text{Test and inspection costs}}{\text{Product units made}}$

(5) Value added base: $\dfrac{\text{Total quality costs}}{\text{Value added}}$

(6) Purchases base: $\dfrac{\text{Incoming inspection costs}}{\text{Value of purchased parts}}$

(7) Quality split: $\dfrac{\text{Prevention costs}}{\text{Total quality costs}}$

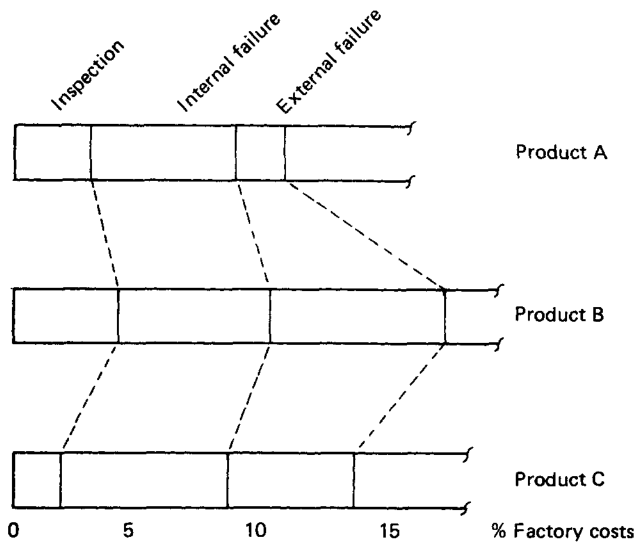


Figure 17.3 Comparative product quality costs

These should be periodically reviewed to reflect any changes which may have occurred in the base used, e.g. increased automation will reduce the direct labour cost, and a change in the product mix may affect the overall blend of quality and/or other costs.

As the organisation becomes skilled in identifying and reporting on quality costs, a more detailed analysis can be made of the breakdown of these quality costs by applying similar ratios to the above but on a departmental and/or product split, as demonstrated in *Figure 17.3*.

Charting techniques, amongst others, can aid in analysis and indicate when action is required.

17.1.4 Quality and the law

There are several areas of civil and criminal law in which quality plays an important part. It can be the ultimate adverse publicity to lose an action under the impartial, and public, gaze of the courts.

17.1.4.1 Civil law

There are three main areas where quality plays an important role.

Vicarious liability This states that an organisation is responsible for the actions of its employees if they occur in the course of that employment.

Contracts In civil law a contract is based on offer and acceptance. The customer may lay down certain specifications, inspections, etc. Terms related to conditions can be expressed in the contract, implied by custom and trade practice or by law, e.g. the Sale of Goods Act (1971) or the Supply of Goods and Services (1982). They include:

(1) goods/services must fit their description,
(2) goods must be fit for the purpose intended, and
(3) goods must be of merchantable quality.

Note that these are *absolute* duties of a seller and the seller is liable even though they have taken all possible care to ensure that these conditions are met.

Quality in all its aspects ensures that the specification can be met before the contract stage and is maintained afterwards. Penalties for failing to fulfil a contract include the customer rescinding the contract and leaving the seller with costs which it cannot recover. In some cases it can lead to penalties being applied in addition.

Consumer Protection Act (1987) This introduces *strict liability* regarding injuries received because of a defective product. A claimant need only prove that:

(1) he, or she, was injured,
(2) the product was faulty, and
(3) the defendant was responsible for the product's circulation.

Note that it is the producer, or importer, who is liable under this act. Negligence need not be proven, nor does *privity of contract*; the plaintiff only needs to show that, on the balance of probabilities, a defect in the product caused the damage.

17.1.4.2 Criminal law

There are several acts which stipulate criminal offences, where again quality plays a part in ensuring no liability occurs.

Trades Description Act (1968) This makes it an offence to make false statements about services or facilities.

Consumer Safety Act (1978) This allows the Government to set regulations about the standards which have to be achieved by specific products, e.g. motorcycle helmets, fabrics for children's nightwear, foam stuffing for furniture.

Health & Safety at Work Act (1974) This lays down an obligation on the part of suppliers to ensure that articles, or substances, be designed and constructed to be safe and without risk to health. Further duties are also laid down to carry out any necessary testing and examination, to carry out necessary research and to provide adequate information for the article's use, storage and transportation.

Consumer Protection Act (1987) This also lays down the General Safety Requirement which states it will be a criminal offence to supply unsafe consumer goods in the UK. In addition, this act consolidates and improves the regulation making power available under previous laws. Many British or international standards will be used as the basis for compliance under these regulations.

Note that in all court action, the court will take into account all relevant circumstances including:

(1) the manner in which the product is marketed;
(2) any instruction, or warning, given with the product;
(3) what reasonably might be expected to be done with the product; and
(4) the time the producer supplied the product.

In all cases, the available state of scientific and technical knowledge will also be taken into account. If it was possible for a defect to be discovered, then ignorance of this fact is seldom acceptable as a defence. It is seldom possible to contract out of these obligations, and an attempt to do so with consumers is in fact unlawful under the Unfair Contracts Act (1977).

In legal matters, the best way to avoid adverse actions is not to commit the offence. Quality is the key to prevention.

17.1.5 Organisation and the duties of a quality-assurance section

Quality assurance must have access to the highest levels within an organisation. Although personnel must have clear responsibility for certain quality tasks, the responsibility also resides in every position within the organisation. Tasks which may require separate personnel are listed below.

Quality engineering
(1) Establishment of quality requirements from the customer's needs.
(2) Analyse performances and recommend improvements.
(3) Assist in setting standards and review criteria.
(4) Determine and produce quality procedures.
(5) Conducts quality audits.
(6) Analysis of customers' complaints.

Process control engineering
(1) Appraise, interpret and implement quality plans.
(2) Conducts drives to reduce quality costs.
(3) Conducts specific tests on equipment for process quality studies.
(4) Calibrates test and measuring equipment.
(5) Records and analyses quality data.
(6) Performs testing, sampling and quality audits.

Information and equipment
(1) Designs, constructs and proves out test and measuring equipment.
(2) Provides calibration schedules for in-process control equipment and test/measuring equipment.

17.2 Quality through integrated design

17.2.1 The role of design with regard to quality

Design is an iterative process which proceeds from establishing the initial market need to selling to the customer. It involves going through a series of phases such as specification, conceptual design, detail design and manufacture.

It is during the specification and conceptual design phases that a product's ability to be fit for its intended purpose is initially realised. *Table 17.1* shows that this ability to meet a given need is determined by a combination of design, material and the processes involved in its manufacture.

It is estimated that, on average, failures during design are responsible for 80% of a product's quality problems. Design failures are also more liable to be passed on to the customer than are manufacturing failures.

Design must be based around the following criteria:

(1) minimum functional requirements,
(2) maximum environmental expectations,
(3) cost limitations, and
(4) safety requirements.

It is design's function to create exact and clear communications in the nature of:

(1) functional specifications,
(2) detail design, and
(3) manufacturing instructions.

Each of these should show clearly the acceptance and rejection criteria and how they have to be tested.

17.2.2 Specifications

The specification is the official means of communicating the needs or intentions of one party to another. It may be written from a using, designing, manufacturing or selling aspect. It should be written in clear unambiguous terms of the optimum rather than maximum or minimum quality grade required to fulfil the customers' need.

Where an appropriate British, or other, standard is already in existence which wholely covers the specification, this should be referred to. Where a standard only partially covers the need, it should be quoted with details of the variation.

Where no standard exists and a specification needs to be drawn up, then BS 7373: 1991 *Guide to the Preparation of Specifications*, gives a comprehensive list of items which may be required in the specification (see *Table 17.2*).

17.2.3 Modifications to design

Over time all designs will be modified. Strict control is essential to ensure that any necessary modifications are made in a cost-effective manner and that all parties concerned are notified and in agreement. It is very important from the quality angle that the correct modification is being referred to in manufacture and parts ordering.

Sometimes, temporary departures from the specification may be requested and these must be especially well controlled. In particular, the following must be fully documented.

(1) *Concession*—the authorisation to use or release a limited quantity of material, components or stores already manufactured but not complying with the specification.
(2) *Production permit*—the authorisation to proceed to manufacture a limited batch using a material or process which does not comply with the specification.

The use of the above should be strictly limited as they can often reflect on the quality image of the organisation.

Table 17.2 Items covered by BS 7373: 1991 *BSI Guide to the Preparation of Specifications*

1.	Title
2.	List of contents
3.	Foreword
4.	Scope of the specification
5.	Role of equipment or material
6.	Definitions
7.	Relevant authorities to be consulted
8.	Related documents and references
9.	Conditions in which the item is to be installed, used, manufactured or stored
10.	Characteristics
11.	Performance
12.	Life
13.	Reliability
14.	Control of quality
15.	Packaging and protection
16.	Information from supplier to user
17.	After-sales service

17.2.4 Process selection

Each process has an inherent range of surface texture and dimensional tolerance which it can produce. The actual range attained being a function of the equipment design, age and the maintenance which it has received. It is imperative, therefore, that the process selected can attain the specified design tolerance required.

Within the possible process range, the cost of production increases substantially as the design tolerance specified tightens. It is therefore important that the widest possible tolerance which can still meet the functional requirements is specified.

Figure 17.4 demonstrates a nominal range of surface roughness values produced by different processes, and *Figure 17.5* demonstrates the relationship of texture to production time. A similar keen relationship exists between processing costs and achieved tolerance (see *Figure 17.6*).

17.2.5 Design techniques to aid quality

17.2.5.1 Product design specification check-list

This is used to produce the specification of what is needed, i.e. What needs must the product be designed to meet? *Table 17.3* lists some of the elements that must be queried to ensure that all aspects of the need are covered.

17.2.5.2 Conceptual design evaluation

Once the specification has been completed, many different ways of evaluating it may present themselves. A weighted evaluation of each concept against the design criteria can be made, as shown in *Table 17.4*.

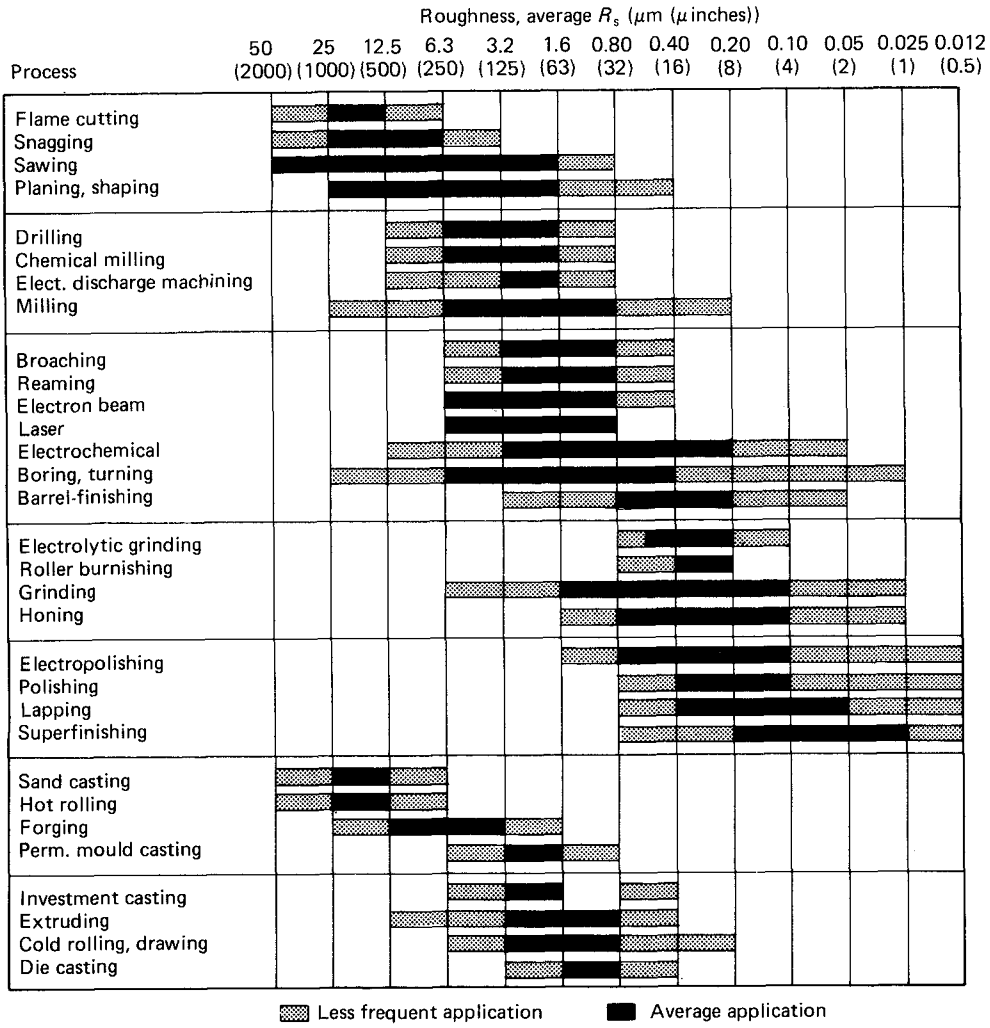


Figure 17.4 The texture capability of processes. The ranges shown are typical of the processes listed. Higher or lower values may be obtained under special conditions

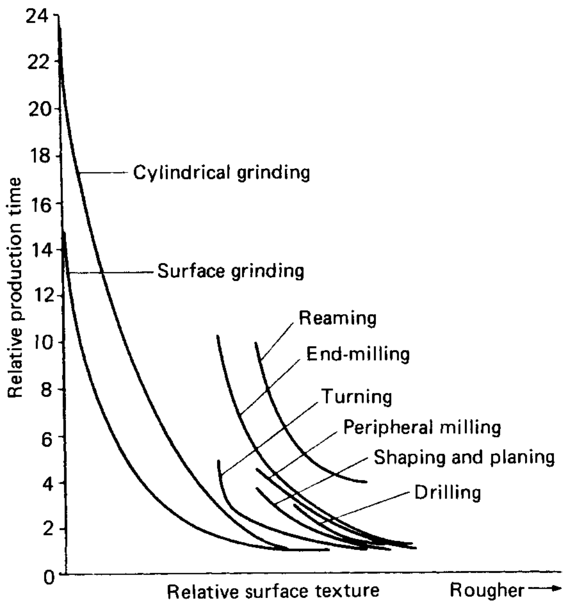


Figure 17.5 Relationship between surface finish and processing time

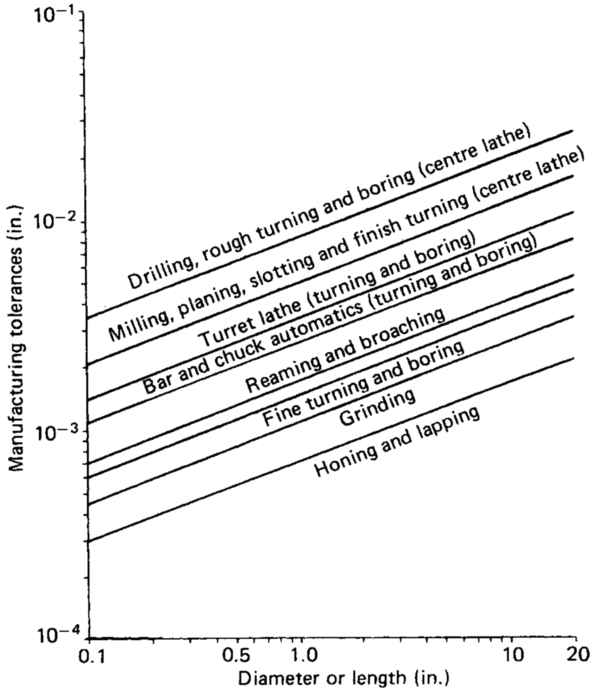


Figure 17.6 Typical tolerances for selected processes

17.2.5.3 Reliability assessment

BS 5760: 1986 *Reliability of Systems, Equipment and Components*, and its three constituent parts gives details of how to go about assessing reliability and managing reliability programmes, and gives many examples of techniques and formulae which can be applied.

Table 17.3 Design specification check-list

Factor	*Points to be answered*
Cost	Is it attainable? Market viability
Producibility	Existing processes. Experience in process. Desirability of new/old process. Supply of parts, material or equipment. Patents
Performance	Speed, frequency, continuity, etc. Noise and vibration level
Strength	Maximum loading. Type of loading (tensile, compression, torsion). Fatigue factors
Size and weight	Restrictions in processing, use or transport
Environment	Ambient conditions of use and storage. Resistance to corrosion
Standards	British or other standard. Testing requirements
Operation	Training, experience and attitude of operating and maintenance personnel
Maintenance	Serviceability. Replaceability. Specialised training and tooling
Life	Overall life. Time between services and overhauls
Reliability	Market need/expectation
Product life	Production span. Time till launch
Competition	Match with existing products
Storage	Method of packaging. Method of shipping. Method of bulk storage. Shelf-life
Safety	Interaction with people. Legal requirements. Materials to be avoided. Special features required. Effect on environment. Recyclability

Reliability can be improved by minimising the parts in use, sticking to known reliable suppliers and components and de-rating an item's operating range, e.g. ensuring a unit rated for 15 A is used in a circuit normally operating at 5 A.

17.2.5.4 Failure mode, effect and criticality analysis

This is one of the techniques referred to in BS 5760: 1986. The aim of this technique is to apply a disciplined, documented approach to evaluating the design of a system, machine, component or process. It involves working through the design

Table 17.4 Design analysis table for a mug washer

Attribute	*Maximum points*	*Design* A	B	C	D	E	F
Dries	10	10	10	7	9	2	10
Cleans insides	10	8	0	6	9	9	0
Cleans outsides	10	8	8	10	10	10	8
Range of mug size	10	10	10	2	8	8	10
Needs services	5	4	3	5	3	4	4
Design complexity	5	4	2	4	2	5	4
No. of components	10	7	3	8	1	9	9
Standard components	10	8	7	8	4	8	9
Ease of maintenance	10	8	8	7	9	5	9
Ease of manufacture	10	8	5	8	6	9	7
Ergonomic soundness	10	10	8	10	10	9	10
Reliability	10	9	7	9	8	8	8
Life expectancy	5	4	2	3	3	4	4
Size	5	5	3	5	4	5	4
Noise	10	10	9	8	6	5	9
Automatic cycle	10	10	10	10	10	7	10
Cycle time	5	4	4	4	5	5	4
Safety in use	10	8	8	7	8	9	8
Approximate cost	5	4	4	3	3	5	3
Total	160	139	111	124	118	126	130

considering in turn the effect and consequence of each failure or malfunction which *could* occur.

On complex systems, it is often necessary to break the main system down into various subsystems. Each subsystem can then be analysed with the result of these being recombined to give the analysis of the main system. The technique used is to complete the analysis form (see *Figure 17.7*) filling in the appropriate data:

Column 1 Description of component.
Column 2 Brief indication of component's function.
Column 3 List *all* possible failure modes of the component.
Column 4 Describe the effect that *each* failure will have on the overall function of the system.
Column 5 List *all* the possible causes of each failure.
Column 6 Indicate the *probability* of each failure cause arising: 1, low probability; 10, certainty.
Column 7 Rank the *severity* of each failure in terms of its effect on the function of the system: 1, only minor; 10, total failure or safety hazard.
Column 8 Rank the difficulty of *detecting* the failure: 1, easily detected; 10, unlikely to be detected.
Column 9 Calculate risk number (=product of columns 6, 7 and 8). This indicates the relative priority of each failure.
Column 10 Give a brief description of corrective action needed, who is responsible, and the expected completion date.

Data on probability can readily be obtained from records of past failures of components and systems coupled with the knowledge of users.

17.2.5.5 Assembly analysis

There are several proprietary systems which analyse an assembly with a view to reducing parts count and simplifying the assembly operations. These have a distinct effect on both quality and costs.

Figure 17.8 shows the loops involved in the design for manufacture procedure developed by Lucas Engineering & Systems.

17.3 Standards

17.3.1 British Standards Institution

Throughout the world each nation has standard institutions to ensure that products are marketable and interchangeable. In Britain this task is overseen by the British Standards Institution (BSI). This is an independent, non-profit-making organisation which receives most of its income directly as a result of supplying its services to Government, industry and commerce. Some of these services are described below.

17.3.1.1 Preparation of British Standards

British Standards are documents which spell out the essential technical requirements for a product, material or process to be fit for its intended purpose. They are prepared, published and up-dated by the BSI in response to the needs of the consumer, industry and the Government.

The use of standards assists in the simplification of manufacturing, stocking, cataloguing and specifying. In many cases the standard itself clarifies basic safety requirements.

Standards influence every stage of the industrial process—materials, dimensions, performance, requirements, codes of practice, test methods and even terminology. They are an essential ingredient for a safe, competitive and cost-effective business.

As one of the world's leading standards organisations, the BSI is active in the promulgation of international standards. This work is carried out through the International Organisation for Standardisation (ISO) and the International Electrotechnical Commission (IEC). Whenever possible, international standards are adopted as British Standards and are given dual ISO/BS or IEC/BS numbers.

17.3.1.2 Information services

The BSI has an extensive enquiry section which can identify any British, other national or international standard, place an order, and arrange translation and delivery. An extensive database is maintained which can be accessed directly by members' computing facilities.

17.3.1.3 Technical help to exporters

When selling goods abroad, knowledge of and compliance with foreign technical requirements is often critical to the success of any venture. This service can identify the appropriate standard for almost any product in most countries and assist in a detailed interpretation of the standard.

17.3.1.4 Assessment and certification

The use of third party certification systems allows manufacturers to demonstrate publicly their commitment to quality. The appearance of an independently awarded certification mark assures the purchasers of the quality of a product. Schemes which the BSI administers or assists in include the following.

Kitemark (Figure 17.9) If the Kitemark appears on a product, it means that the product has been independently tested

Product – I.C. engine	Subsystem – in line F.I. pump	*P* Probability of occurrence		F M E C A No. PR001
Process –	Location –	*C* Severity of failure		Date
System –	User –	*D* Difficulty of detection		Engineer
		R.N. Risk Number = $P \times C \times D$		

(1) Part No. Component	(2) Function	(3) Failure mode	(4) Effect of failure	(5) Cause of failure	(6) *P*	(7) *C*	(8) *D*	(9) *R.N.*	(10) How can failure be eliminated or reduced
Cams and follower rollers	Plunger drive	Fatigue pitting	Loss of power irregularity (or complete functional failure in extreme cases)	Overload (e.g. nozzle blockage, dirty or no oil, poor surface finish)	2	3	10	60	Supplier to determine failure definition. Limit operating pressure
Centre bearing bridge	Support camshaft centre bearing. Ties pump sides at centre	Cracked across centre of bridge	Mainly cosmetic. Slight oil seepage		1	2	10	20	Problem exists on present die cast pumps. Supplier likely to shell mould on our B cylinder pump therefore more material on bridge. To be assessed when samples are tested
Delivery valve holder		Erosion in bore	Nozzle blockage, loss of performance	High flow rates. Cavitation	5	3	10	150	To be assessed when samples are tested
Delivery valve	Unloading	Consistency U/L wear	Engine performance pump phase and delivery balance	Manufacture, adjustment poor filter/water	3	9	4	108	M/C test. L.P. systems detail
Delivery valve spring	Controls unloading	Fatigue/erosion/ wear (omission)	Loss of engine performance, smoke	Hydraulic duty	4	4	4	64	Specification detail
Delivery stop peg	Control delivery lift	Wear—crushing (omission)	As above	As above	2	4	3	24	Injection rates
Plunger element		'F' slot polishing	Seizure or sticking leading to functional failure	Particulates produced by erosion of pump body in transfer gallery or erosion shield	1	10	10	100	Slots now roller into element prior to hardeing and grinding. To be assessed when samples are tested

Figure 17.7 Failure mode, effect and criticality analysis. (Courtesy of British Standards Institution)

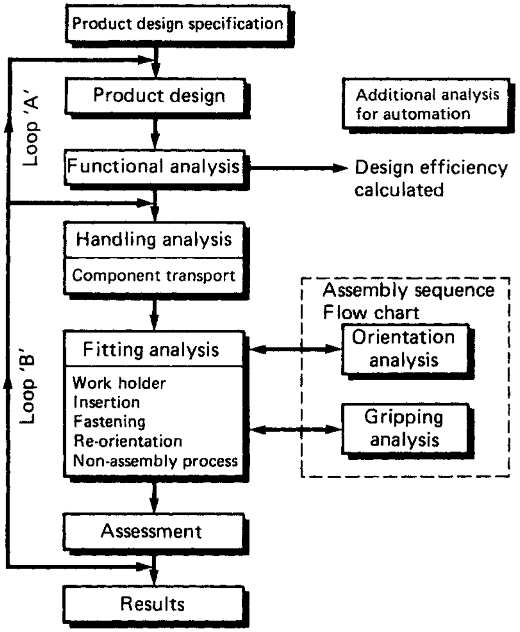


Figure 17.8 Design for assembly procedures. (Courtesy of Lucas Engineering & Systems Ltd)

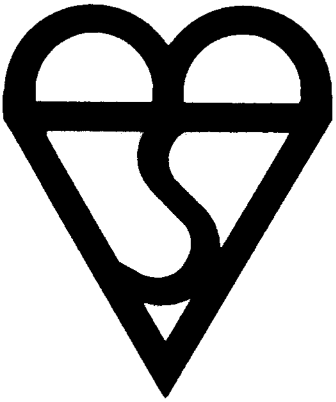


Figure 17.9 The BSI Kitemark

against the relevant British Standard and that the company's quality system has been assessed against BS 5750: 1987. When a manufacturer has been licensed by the BSI, unannounced visits are made to the factory by the BSI Inspectorate to ensure that the agreed quality levels are being maintained.

Safety mark (Figure 17.10) Similar procedures to the Kitemark apply, but with specific concern for safety in the product's normal use.

Registered firm of assessed capability If a company conforms to the relevant part of BS 5750: 1987/ISO 9000, then they may be licensed to use the Registered Firm symbol (see *Figure*

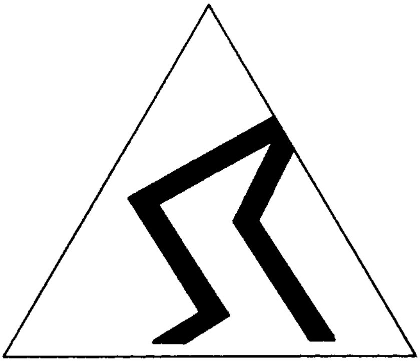


Figure 17.10 The BSI safety mark

Figure 17.11 The BSI Registered Firm of Assessed Capability mark

17.11). For electrical components, the certification scheme is BS 9000: 1989.

17.3.1.5 BSI inspectorate

In addition to the assessment and routine surveillance for the BSI certification schemes, the BSI Inspectorate can also cater for the needs of an individual company or government body. This could entail assessment of a potential supplier, continually auditing quality levels, or ensuring that imported/exported goods are of the correct quality before despatch. This could be on a one-off basis or continuously.

17.3.1.6 Testing services

The BSI Testing Service carries out a portion of the testing required as part of the certification schemes, as part of its duties. Tests can also be carried out on a wide range of products against a variety of national and international standards, government regulations or company specifications.

17.3.2 Quality systems (BS 5750/ISO 9000)

Standards are the base point for quality levels and many standards are produced for quality techniques, and even for an organisation's quality system itself. To become registered under BS 5750: 1987/ISO 9000 has become very important, as this states that a company has the basic procedures and discipline to ensure that their output meets the customers' requirements.

In 1987, BSI revised the 1979 version of BS 5750 to become dual-numbered with ISO 9000, so that registered firms could deal on an international scale. The revised structure is:

Part 0: Section 0.1 *Guide to Selection and Use*
Part 0: Section 0.2 *Guide to Quality Management and Quality Systems Elements*
Part 1: *Specification for Design/Development, Production, Installation and Servicing*
Part 2: *Specification for Production and Installation*
Part 3: *Specification for Final Inspection and Test*

As can be seen in *Figure 17.12*, the coverage of the various parts is extensive and permeates all aspects of a company's organisation. It is worthwhile examining each clause in part 1 in a little more detail:

4.1 The management responsibility for maintaining quality must be clearly identified. This includes allocation of suitable personnel, the organising of quality audits and periodic review of the effectiveness of the quality system itself.
4.2 The quality system itself must be fully planned, documented and maintained. It should cover all aspects of the organisation including staff, equipment and records.
4.3 The contract with the customer sets out the specification that the organisation has to supply against. It must be carefully checked to ensure it is clear, unambiguous and capable of being met by the available resources.
4.4 Design control is needed to ensure that the design meets the customers' needs. This includes organising necessary interfaces and keeping records of design assumptions and calculations.
4.5 All procedures for changes/modifications must be especially tightly coordinated and authorised. Especially important is the procedure for ensuring the correct document is in place when required.
4.6 Correct quality depends on correct procurement procedures so that the items bought in meet the requirements. This involves vendor assessment, clear purchase specifications, incoming inspection and, of course, records.
4.7 Free issue material from the customer must be protected in use and storage when within the organisation's control.
4.8 Material control and traceability is necessary to ensure that all non-conforming material can be traced from suppliers, through the processing stages to customers. This is especially true in safety matters as it may give an opportunity to recover claims.
4.9 All manufacturing should take place to clear, proven instructions. The instructions should emphasise the control of any specialised processes and detail the necessary recording media.
4.10 All testing and inspection prior to, during and after manufacturing must be recorded to confirm compliance with specification. The techniques and equipment used must be clearly specified and their accuracy established. This should include certificates of compliance received.
4.11 Inspection, measuring and testing equipment must be capable of meeting the required precision consistently. Procedures should exist to control and maintain equipment, including scheduled calibration and traceability to recognised measurement standards.
4.12 The inspection status of all material must be clearly identified during all stages. The responsibility for releasing conforming material should be clearly established.
4.13 Non-conforming material must be clearly segregated and identified, including the reason. The subsequent disposition of this must be to clear procedures and the deciding authority recorded.
4.14 Corrective action should follow lapses in conformity as it is only the identification of causes that ensures subsequent conformity economically. This extends to vendors and subcontractors.
4.15 Handling, storage and transportation are all part of the supply train. Handling procedures should detail product protection to prevent damage and/or contamination which affects the products' usefulness.
4.16 Records are evidence of what occurs during any manufacturing and must contain all instructions, actions and results in an accessible system.
4.17 Auditing the quality system ensures it operates efficiently. Where improvements are identified, the responsibility for carrying them out must also be stated.
4.18 Proper training is the foundation of quality. Needs must be correctly identified and training must be effective, preferably by demonstrated records of competence reached.
4.19 Servicing quality must be handled in a similar manner to internal quality matters, including procedures for verification and records.
4.20 Statistical techniques, where used, must be specified, including any associated risk levels.

An organisation wishing to install a quality-control system such as BS 5750: 1987 will find it a useful marketing tool. In some cases, such as when supplying government agencies, it may be a prerequisite for tendering.

17.4 Material and process control

17.4.1 Incoming material

In addition to design and the processes involved, the other determinant of quality referred to in *Figure 17.1* is material. Section 17.6 details the sort of acceptance sampling schemes that can be applied to batches of incoming material. However, the most important method of ensuring one has suitable material available is to carry out a complete assessment of all suppliers' abilities to supply the correct material.

17.4.1.1 Vendor assessment

This assessment will involve visits to suppliers to examine their processing facilities and audit their quality procedures. Moreover, when examining suppliers, one should take into account all the factors which are of importance to the company:

(1) quality, i.e. the level of rejected batches;
(2) price per item, relative to other suppliers;
(3) delivery, i.e. deliveries made on time;
(4) service, i.e. the amount of back-up and flexibility demonstrated.

Vendor assessment is a simple rating of different suppliers against each other under headings so that their overall value to the company can be assessed. It is a matter of selecting the particular factors, such as those above, which are considered important, and applying a weighting to them in relation to their agreed impact on the company:

Clause (or subclause) No. in BS 5750: Part 0. 0.2/ISO 9004	*Title*	*Corresponding clause (or subclause) Nos in*		
		BS 5750: Part 1/ ISO 9001	*BS 5750: Part 2/ ISO 9002*	*BS 5750: Part 3/ ISO 9003*
4	Management responsibility	4.1 ●	4.1 ◐	4.1 ○
5	Quality system principles	4.2 ●	4.2 ●	4.2 ◐
5.4	Auditing the quality system (internal)	4.17 ●	4.16 ◐	—
6	Economics—quality-related cost considerations	—	—	—
7	Quality in marketing (contract review)	4.3 ●	4.3 ●	—
8	Quality in specification and design (design control)	4.4 ●	—	—
9	Quality in procurement (purchasing)	4.6 ●	4.5 ●	—
10	Quality in production (process control)	4.9 ●	4.8 ●	—
11	Control of production	4.9 ●	4.8 ●	—
11.2	Material control and traceability (product identification and traceability)	4.8 ●	4.7 ●	4.4 ◐
11.7	Control of verification status (inspection and test status)	4.12 ●	4.11 ●	4.7 ◐
12	Product verification (inspection and testing)	4.10 ●	4.9 ●	4.5 ◐
13	Control of measuring and test equipment (inspection, measuring and test equipment)	4.11 ●	4.10 ●	4.6 ◐
14	Non-conformity (control of nonconforming product)	4.13 ●	4.12 ●	4.8 ◐
15	Corrective action	4.14 ●	4.13 ●	—
16	Handling and post-production functions (handling, storage, packaging and delivery)	4.15 ●	4.14 ●	4.9 ◐
16.2	After-sales servicing	4.19 ●	—	—
17	Quality documentation and records (document control)	4.5 ●	4.4 ●	4.3 ◐
17.3	Quality records	4.16 ●	4.15 ●	4.10 ◐
18	Personnel (training)	4.18 ●	4.17 ◐	4.11 ○
19	Product safety and liability	—	—	—
20	Use of statistical methods (statistical techniques)	4.20 ●	4.18 ●	4.12 ◐
—	Purchaser supplied product	4.7 ●	4.6 ●	—

● Full requirement.
◐ Less stringent than BS 5750: Part 1/ISO 9001.
○ Less stringent than BS 5750: Part 2/ISO 9002.
— Element not present.

Notes
1. The clause (or subclause) titles quoted in the table above have been taken from BS 5750: Part 0. Section 0.2/ISO 9004, the titles given in parentheses have been taken from the corresponding clauses and subclauses in BS 5750: Part 1/ISO 9001, BS 5750: Part 2/ISO 9002 and BS 5750: Part 3/ISO 9003.
2. Attention is drawn to the fact that the quality system element requirements in BS 5750: Part 1/ISO 9001, BS 5750: Part 2/ISO 9002 and BS 5750: Part 3/ISO 9003 are in many cases, but not in every case, identical.

Figure 17.12 List of quality system elements per part of BS 5750: 1987. (Courtesy of British Standards Institution)

Maximum points for quality 40
Maximum points for price 25
Maximum points for delivery 20
Maximum points for service 15

Each supplier can then be compared against these factors.

Example A supplier has the following profile:

90% of his deliveries are made on time.
The price is 105 p per item against the cheapest quote of 90 p.
95% of delivered batches are accepted on inspection.
His reactions meet our requirements on 80% of occasions.

This supplier would then be rated as:

Delivery rating = 90% of 20	= 18
Price rating = (90 p ÷ 105 p) × 25	= 21
Quality rating = 95% of 40	= 38
Service rating = 80% of 15	= 12
Total rating	89

This supplier's total rating can then be compared with those of all other suppliers for the same material or components, and it can be decided which supplier best serves the company's needs. The total rating can also be used in conjunction with selected suppliers to show where their total service is behind their competitors in order to drive up the general rating of all suppliers combined.

17.4.1.2 Vetted suppliers

Once a supplier has been selected to be the main, or sole, supplier, and we are convinced that the quality of their supply consistently meets our requirements, it should be possible to reduce the need to inspect their incoming goods.

It will still be necessary to monitor their quality, partially by means of occasional quality audits and partially from records of their material/components service within the plant and in service with our customers. To ensure that this can take place it is important to ensure that full traceability of all materials used is possible and maintained.

17.4.2 In-process control

In order to achieve control in any situation, we must know what we wish to produce and how the process cycle goes about achieving this. Ideally, we wish to set up a closed-loop control cycle as shown in *Figure 17.13*. In order to do so we need to be able to specify and quantify all the various aspects. Other sections deal with the process of measuring and comparing against standards and the adjustments that need to be made. This section discusses the inspection points at which measurement can take place.

In any feedback situation, two matters are important:

(1) the *time* that has elapsed since the process was carried out, and
(2) the *quantity* of produce that has been produced meanwhile.

With process control we wish to minimise both these factors.

17.4.2.1 Preprocess inspection

If we inspect the incoming material, as in *Figure 17.14*, we can decide on what process steps are required in order to reach specification. In the extreme, we can decide not to advance material for processing.

A common method is to incorporate a go/no go check on features and sizes into tooling fixtures so that material which is outside the incoming specification cannot be mounted prior to processing.

17.4.2.2 First-off inspection

When a process is first set up, it is important that the initial output is closely monitored to ensure it meets the specification on an ongoing basis.

17.4.2.3 Last-off inspection

When a batch has been completed, special care should be taken with the measurement of the final few items. This is to ensure that the fixtures and tooling are of sufficient standard to be used again. Where the output shows signs of approach-

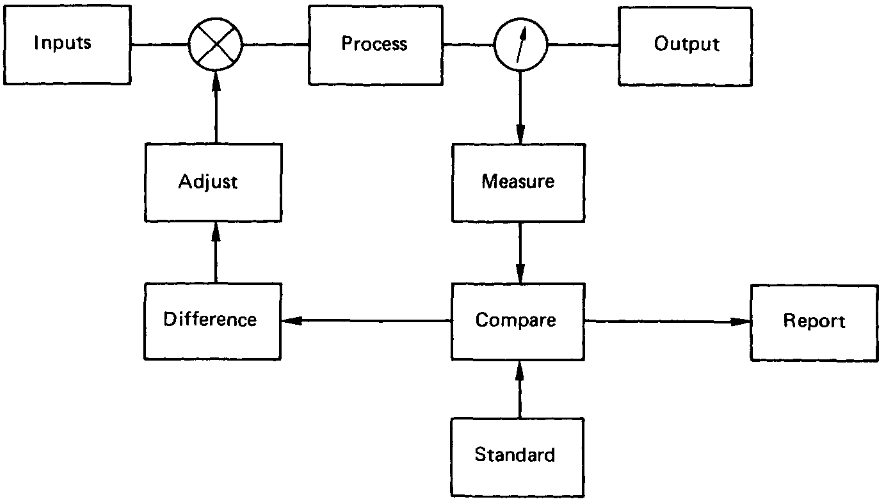


Figure 17.13 Closed-loop control diagram

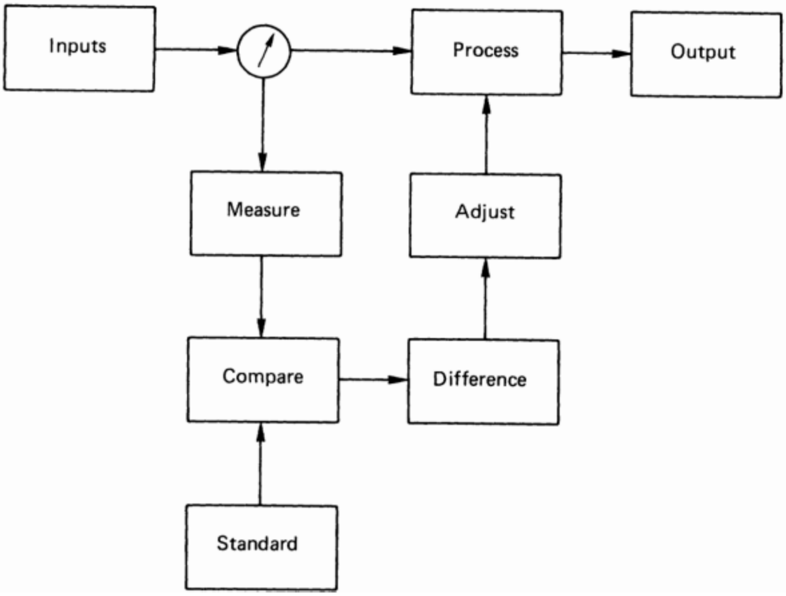


Figure 17.14 Preprocess control

ing the specification limits, then regenerative action on the equipment can be put in hand so that it is ready for a later run.

17.4.2.4 Post-process inspection

This is the normal stage at which the quality of the process is monitored. The location of this should be as close as possible to the process to minimise the opportunity for out-of-specification material to build up.

The first factor to consider is the frequency of post-processing inspection. How often need we carry out the inspection—after every item, once every 100 items, every 15 min, or when? This should be decided after examining the probability of the process going out of control, the integral drift that occurs, e.g. tool wear, and the costs involved both in inspection and in having out-of-specification items.

The second factor is the physical location of the inspection point. Ideally, the operative himself should be able to carry out, or oversee, the inspection process. Therefore, the inspection should take place close to the processing point to allow the operative access to the inspection equipment. If the inspection takes place at a remote station, e.g. a coordinate measuring machine, then the results need to be fed back to the process and/or the operative.

17.4.2.5 Adaptive control

Ideally any defect should be discovered as soon as it has occurred and corrective action initiated automatically. For this to happen adaptive control systems are needed that measure the process variable as it is being produced, or immediately afterwards.

The former can be achieved with in-process gauging, with further processing being determined by measuring the variation remaining until the specification is reached. *Figure 17.15* shows a typical electronic grinding gauge used in such a system.

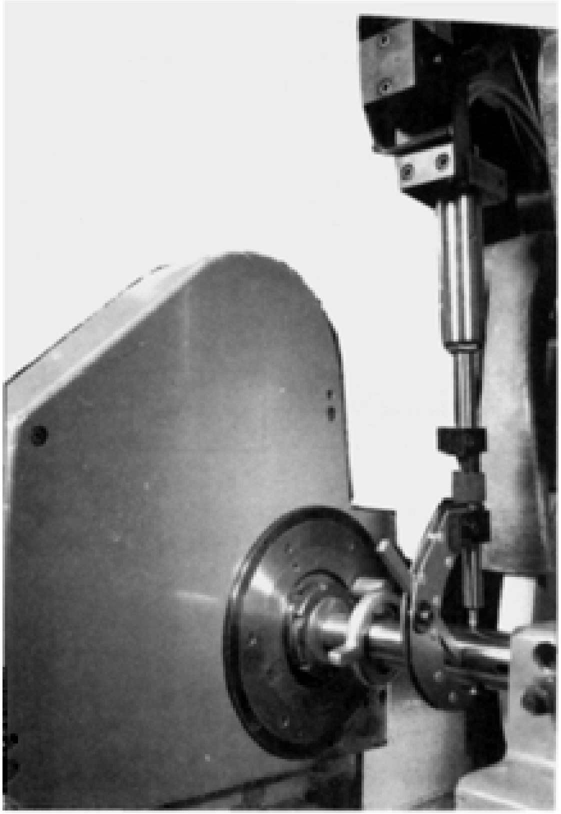


Figure 17.15 An electronic grinding gauge. (Courtesy of Vernon Gauging Systems Ltd)

The latter can be carried out by probes mounted in the processing equipment itself. *Figure 17.61* shows such a probe mounted on a computer numerical control (CNC) machining centre.

The advantage of measuring in the processing location is in the timeliness and ease of feedback. There are disadvantages, however, in that specialised measuring equipment has to be duplicated, processing may have to be interrupted and the environment is less controlled, e.g. residual heat and stresses may be in the product.

17.5 Process capability

The most important determinant of the quality output of a process is the inherent capability of that process to produce in a consistent manner. Determining the range produced by a process and the movement, or drift, of that range is therefore important. These factors will determine if the process will consistently produce within specification or with a degree of substandard output. It also lays the basis for process improvement by identifying exactly where improvements may be possible, or even necessary.

17.5.1 Process variability

It is a relatively easy matter to take continuous measurements of the output of any process. We can then examine what happens to this value over a period of time. This can be examined in two main ways.

(1) *Dispersion* How does the total of the individual measurements vary over the whole of a production run? Does it vary between machines, between shifts or even at different times of the day? We can examine this by plotting a frequency diagram of the particular sizes being produced.
(2) *Drift* How do the individual measurements change as the process continues to operate? We can examine this by plotting individual measurements over time.

17.5.1.1 Dispersion patterns

The easiest method of quickly drawing a rough of the frequency distribution is to produce a tally check (see *Table 17.5*). This is only a matter of noting, by a cross, against each size when that size is produced, and gradually a histogram will be formed. The use of tally checks, however, is limited in the amount of data it can deal with.

Table 17.5 Tally chart of weights from a process

Weight (g)	*Occurrences*
220	XX
221	X
222	XXXX
223	XXXXXXX
224	XXXX
225	XXXXXXXXX
226	XXXXX
227	XX
228	XXX
229	X

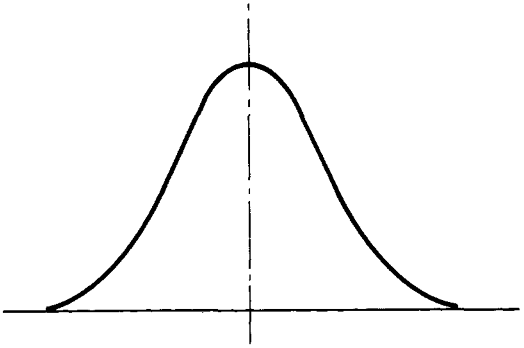


Figure 17.16 The normal distribution curve

It is more meaningful to collect a large amount of data over a significant period of time before forming a histogram or a frequency diagram for analysis. If the measurements vary in a purely random manner, the diagram produced will be the normal, bell-shaped distribution curve (*Figure 17.16*). This dispersion determines the capability of the process consistently to produce items within specification.

It is when the dispersion is set against the specification tolerances that the true capability is seen. When the dispersion is too great (see *Figure 17.17(a)*) then some of the produce from the process *must* be out of specification. This process is basically unsuitable for the present tolerances, unless something is done to bring them into a better match, e.g. widening the tolerance or improving the process. When the dispersion is much smaller (see *Figure 17.17(b)*) then the process can easily produce within the specification. If, however, the dispersion is substantially smaller, the process could be said to be 'too good', and is probably more expensive than necessary. If the dispersion is exactly the same as the tolerances (see *Figure 17.17(c)*) this is ideal. However, great care is required to ensure that the setting remains in line with the tolerances.

Ideally, we need a dispersion between *Figures 17.17(b)* and *17.17(c)*. It needs to be less than the tolerance limits to allow for some drift for unavoidable events such as tool wear, but not substantially so.

When the dispersion does not form a true normal curve, there is normally something wrong with the process itself. A departure from a random distribution is usually due to a cause. This shows up in the dispersion diagram as a bias, or as items occurring outside the normal curve. Investigation, if able to identify the causes, can then eliminate them.

17.5.1.2 Drift patterns

In addition to the range produced by a process, the way the output varies over time is also important. This can be seen by plotting the characteristic against a time base (see *Figure 17.18*). Again, it is when this plot includes tolerance zones that the true effect of any change is revealed.

Where the change is random, the range is clearly seen over a short period. A broad scatter (see *Figure 17.19*) means that care is required to ensure that the outermost items remain within tolerance. A tight band shows that less care is required in setting and less frequent adjustments are needed.

Where a gradual but distinct change occurs, i.e. drift, this can be used to determine the time interval between inspections. It also shows that initial settings can be made towards one of the tolerances so that the drift allows the maximum

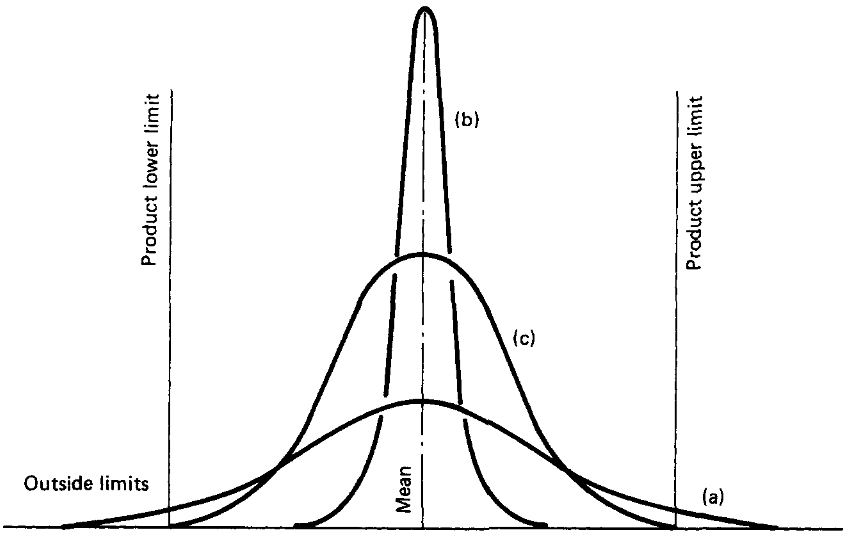


Figure 17.17 Comparison between process variability and product specification limits

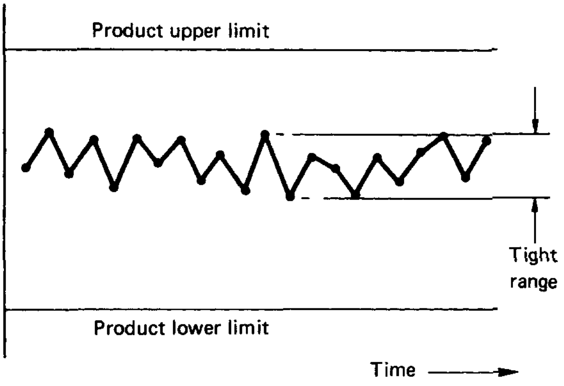


Figure 17.18 Component measurement over time—low variability

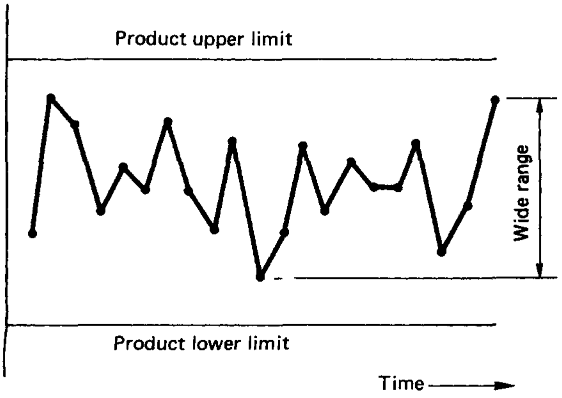


Figure 17.19 Component measurement over time—high variability

time before items exceed the other tolerance line (see *Figure 17.20*).

17.5.2 Problem identification and highlighting techniques

17.5.2.1 Pareto curves

It is important that effort in quality improvement be targeted to achieve the most beneficial results. Pareto charts effectively show up the relative frequency of particular types of faults occurring. If faults are collated by cause, the results can be charted on a histogram, with the heaviest occurrences at one end (see *Figure 17.21*). These can be converted into percentages of all faults for a fuller effect. Normally, from this analysis it can be seen that the majority of the faults occur in relatively few areas. It is in these areas that effort should be directed to show the largest quality improvements.

17.5.2.2 Fishbone diagrams

As there are many possible causes for variations in quality, it is often useful to carry out a brainstorming session to identify possible causes. During these sessions, it is helpful to produce a fishbone, or Ishikawa, diagram to show all possible causes. The effect being investigated is shown at the end of a horizontal arrow (see *Figure 17.22*). Potential causes are then shown as labelled arrows entering this horizontal arrow. Each main arrow may have other arrows entering it as possible subcauses. The subcauses may, in turn, be due to a number of possible sub-subcauses.

The process is continued, perhaps using a series of fishbone diagrams at different levels, until all possible causes, however unlikely, are listed. Once this has been done, each cause can be considered for further detailed investigation.

17.5.2.3 Cusum charts

The cusum (cumulative sum) chart can detect trends in any measured value very quickly. It is particularly useful for

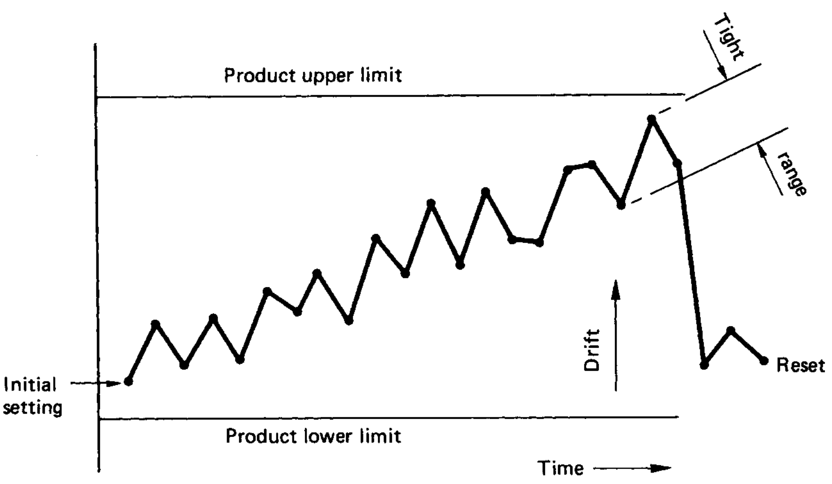


Figure 17.20 Initial setting when drift occurs

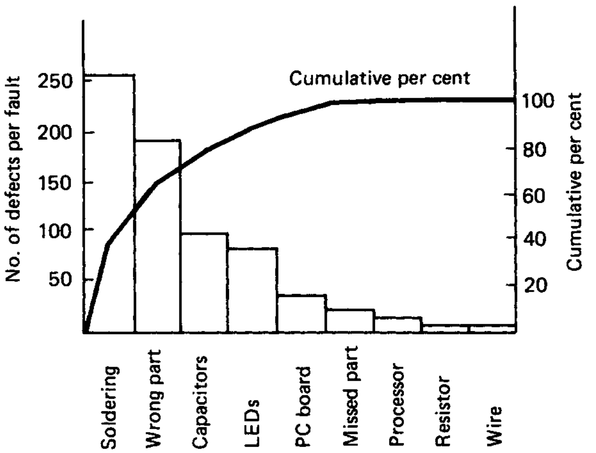


Figure 17.21 Pareto curve for electronic faults

plotting the evolution of processes because it presents data in a way that enables the eye to separate true trends from random variations. Cusum charts will detect small changes in mean values quickly and can be used for variables or attributes. In essence, a reference or target value is subtracted from each successive sample observation and this result accumulated with the previous ones. The cumulative value is plotted and trend lines can readily be seen on the resultant graph.

If the trend line is approximately horizontal, the average value of the sample is the same as the reference value. A downward slope shows that the average being achieved is less than the reference value; an upward slope shows that the average is greater. Where the slope is constant, the difference is also constant. This means that, although the average is different, it remains at a constant value. Hence only resetting is required to align the average with the reference value. When the slope is changing it shows the trend towards, or away from the reference value.

An example of some data, the resultant cusum and the actual-size control chart are shown in *Figure 17.23*.

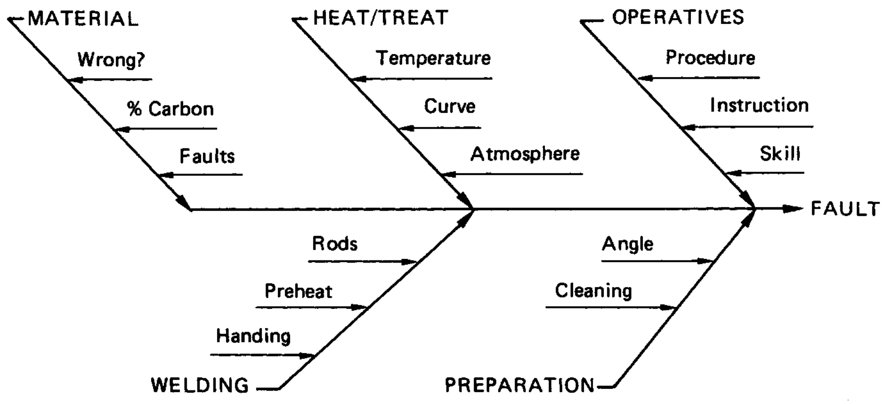


Figure 17.22 Fishbone diagram

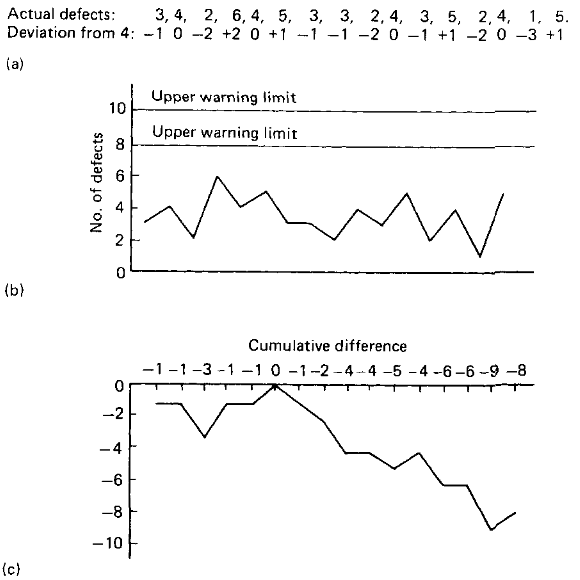


Figure 17.23 Example of a cusum chart. (a) The expected number of defects in each batch of 2000 items was 4. The actual number of defects in successive batches is shown. The cusum technique enables us quickly to determine if the present level is better or worse than expected. This gives rise to a number-of-defects chart (b) based on an expected value of 4 defects. It is difficult to see from (b) if the process is working at the expected average rate of defects or not. However, when the data are plotted as cumulative differences (c) it can readily be seen that the process is performing slightly better than expected, as in this particular case a negative change in the slope of the curve is good. The cumulative average is 3.5 defects as compared with the 4 allowed. The slope of the line in (c) shows the change in the deviation on a constant basis

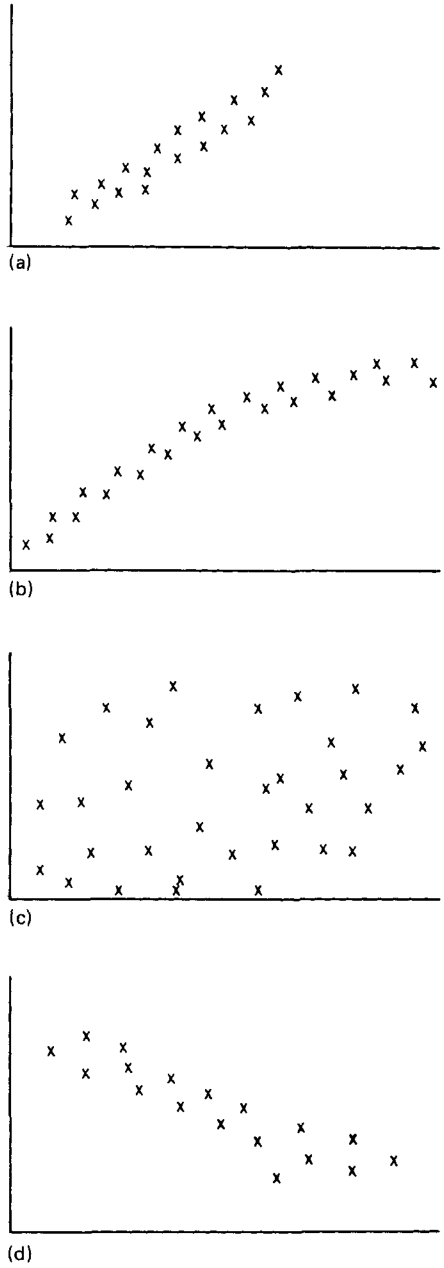


Figure 17.24 Correlation patterns: (a) linear, positive; (b) varying ratio of correlation; (c) negligible correlation; (d) negative correlation

17.5.2.4 Correlation chart

Where there is thought to be a connection with a particular input variable and the quality level, it is beneficial to plot them together to determine if there is any correlation between them. *Figure 17.24* show various correlation patterns which could be apparent.

17.5.2.5 Measles charts

In complex assemblies, it is often useful to mark a sketch of the assembly with the location of any fault which arises. A clustering of marks (see *Figure 17.25*) readily identifies the location of recurring problems. This technique is particularly useful for electronic assemblies, but can also be used in pressure circuits and casting and moulding operations where porosity and cracking can occur.

17.6 Product acceptance sampling schemes

17.6.1 Product inspection frequency

It might be thought that the only way to ensure a batch of products entering, or leaving, a facility is made to specification is manually to inspect every attribute of each item, i.e. to carry out 100% inspection. However, it has been proven that, especially in the case of repetitive production, 100% manual inspection often does *not* in fact give this assurance.

The reasons for the failure of 100% manual inspection to achieve its aim centre around normal human error in measuring and interpreting results, coupled with the attendant monotony associated with maintaining alertness when carrying out

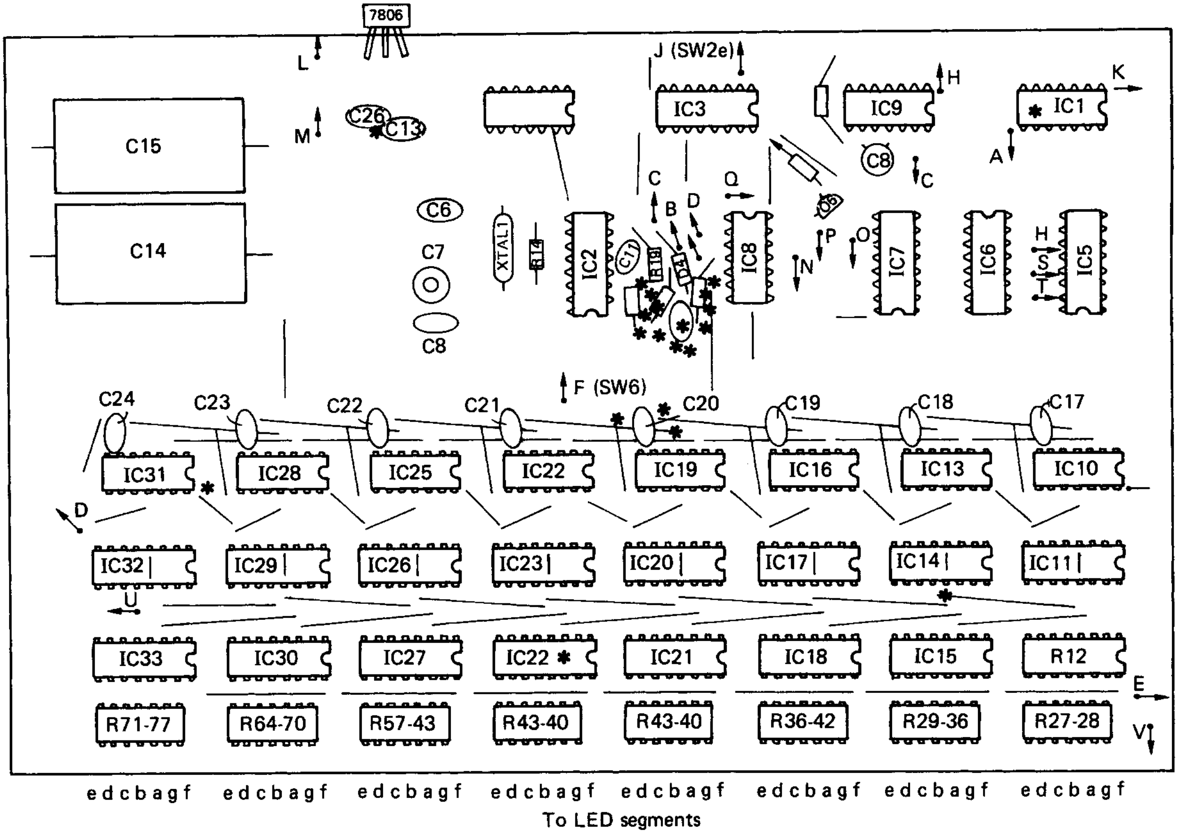


Figure 17.25 A measles chart of an assembly. * Denotes a fault

repetitive simple measuring tasks. In addition, measuring all the attributes of each item is a long and costly process.

In some cases it is impossible to test every item for a particular characteristic, e.g. ultimate strength, because the only available method for testing that characteristic involves testing to destruction.

In many industries it is common to determine the characteristics of the whole population by examining a relatively small sample, e.g. opinion polls are used to gauge the public's reaction to political events. It is therefore equally valid that taking a random sample from a batch of items and measuring pertinent attributes will give an impression of what the remaining, unmeasured items are like. This technique involves selecting a suitable sample size dependent on the batch size. Using probability theory and the normal quality level, the expected number of defectives in the sample can be calculated.

If the inspection of the sample produces less defectives than expected, then the whole batch is probably within the normal quality level and therefore can be accepted; conversely, if the number of defectives found is greater, then the whole batch is rejected, downgraded or sent for 100% screening.

17.6.2 Acceptance sampling schemes

BS 6001: 1991 *Sampling Procedures and Tables for Inspection by Attributes* details the following sampling schemes. The examples shown against each sampling scheme are all based on a batch of 1000 items with a normal quality level of 1% defectives being subjected to normal inspection.

17.6.2.1 Single sampling

This involves taking from a batch a single, representative sample (size n) and testing it for defects. The lot is accepted if the sample contains Ac or less defective items. The lot is rejected if it contains Re or more rejects. Normally, $Re = Ac + 1$.

Example

Batch lot size = 1000.
Sample size n = 80.
Accept lot if defectives in sample $Ac \leqslant 2$.
Reject lot if defectives in sample $Re \geqslant 3$.

17.6.2.2 Double sampling

Under this plan, a smaller sampler is taken initially than in single sampling. If the quality is such that the batch is either significantly good, or bad, then a decision to accept or reject the lot can be taken on the results from this initial sample. If the evidence is inconclusive, then a further sample (normally the same size as the first) is drawn and tested.

Example

Batch lot size = 1000.
First sample size $n_1 = 50$.
Accept lot if defectives in sample $Ac_1 = 0$.
Reject lot if defectives in sample $Re_1 \geqslant 3$.

If defectives in first sample is 1 or 2, then take second sample of size $n_2 = 50$.
Accept lot if cumulative defectives from both samples $Ac_2 \leqslant 3$.
Reject lot if cumulative defectives from both samples $Re_2 \geqslant 4$.

17.6.2.3 *Multiple sampling*

The principle here is similar to double sampling, except that more than two successive samples are allowed for before a decision is made. The samples are small and BS 6001: 1991 allows for up to seven samples to be drawn if the earlier ones are inconclusive.

Example

Batch lot size = 1000.
First sample size $n_1 = 20$.
Lot cannot be accepted on basis of first sample, even if no defectives found.
Reject lot if defectives in sample $Re_1 \geqslant 2$.

If defectives in first sample are less than 2, then take a second sample of size $n_2 = 20$.
Accept lot if cumulative defectives from both samples $Ac_2 = 0$ (zero).
Reject lot if cumulative defectives from both samples $Re_2 \geqslant 3$.

If cumulative defectives from first two samples is 1 or 2, then a third sample ($n_3 = 20$) is taken and a similar choice is made whether to accept the lot, reject it or continue sampling and testing.

17.6.2.4 *Sequential sampling*

Under this plan there are no fixed sample sizes. The result of testing each sequential individual item is accumulated until sufficient evidence exists to accept or reject the batch. The scheme does allow for a maximum cumulative number of tests at which the acceptance and rejection numbers under a multiple-sampling plan would apply.

The procedure up to this point involves the use of two factors—the handicap H, and the penalty p—both of which are extracted from the relevant tables in BS 6001: 1991. A continuous score is kept using the formula:

$$\text{Score} = (H + \text{Number passed}) - (p \times \text{Number failed}) \quad (17.1)$$

The testing continues until either:

(1) the score reaches $2 \times H$, in which case the lot is accepted; or
(2) the score falls to zero, in which case the lot is rejected; or
(3) the maximum sample size is reached, in which case the multiple-sampling-plan criteria are applied.

This procedure is too complex to be carried out in a normal working environment. Tables showing break points are therefore issued where these schemes are in operation.

Example

Batch lot size = 1000.
Handicap $H = 40$.
Penalty $p = 32$.
Maximum sample size $n = 140$.
If maximum sample size is reached then:
accept lot if accumulative defectives $Ac \leqslant 4$;
reject lot if accumulative defectives $Re \geqslant 5$.

17.6.3 Factors affecting the choice of sampling plan

17.6.3.1 *Simplicity*

Single sampling is obviously the easiest to understand and administer. As the number of samples possibly required increases, so does the complexity.

17.6.3.2 *Ease of drawing sample units*

Where drawing samples is relatively easy, there is little difference between using the different plans. If drawing samples is difficult then it may be better to either use the single sampling plan, or to draw out the maximum size sample possibly required and select out the intermediate samples from this.

17.6.3.3 *Duration of test*

If a test is lengthy yet can be applied to a number of sample units simultaneously, then it is often better to do so. If, on the other hand, only a small number of items can be tested at a time, then multiple or sequential sampling may be best.

17.6.3.4 *Multiple attributes*

The more complex the item in terms of number and classification of possible defects, the more simple the scheme desirable.

17.6.3.5 *Cost of test*

Sequential sampling can normally give rise to a decision on a smaller number of tests than can other plans. This may be particularly important if the test is to destruction.

17.6.4 Use of BS 6001 sampling procedures and tables

Having selected what type of sample plan to use, e.g. single sample, it is a simple matter to extract the pertinent details from the tables within BS 6001: 1991.

17.6.4.1 *Inspection level*

BS 6001: 1991 allows for seven different inspection levels, which are listed in *Table A17.1* in the Appendix to this chapter. Unless otherwise specified, inspection level II should be used. Level I may be specified where less discrimination is required, or level III if greater discrimination is required. The other four levels (S-1 to S-4) which are given may only be used in special circumstances where very small sample sizes are required, but these levels give a greater risk of producing misleading results.

17.6.4.2 *Code letter*

The sample size is designated by a code letter (see Table A17.1 in the Appendix to this chapter). The code letter is selected from the combination of batch size and inspection level. Therefore, for the previous example, with batch size of 1000 items the code letter to be used is J for inspection level II.

17.6.4.3 Acceptable quality level

The acceptable quality level is that level of defectives which can be considered satisfactory as a process average. At this level the sampling plan should select the majority of the lots from a process, assuming that this average is being maintained. The acceptable quality level is used with the code letter to determine the characteristics of the desired sampling plan.

17.6.4.4 Detail of sample plan

The code letter and the acceptable quality level are used in BS 6001 : 1991 to obtain the details of the sampling plan. BS 6001: 1991 contains three classes of inspection (normal, tightened and reduced) for each of the sampling schemes mentioned above.

The procedure is to turn to the table for the selected scheme and class and cross match the code letter against the acceptable quality levels listed. *Tables A17.2* and *A17.3* in the Appendix to this chapter are used to select a single and double normal sample plan for the example batch of 1000 items, respectively.

Initially one reads down the first column until the code letter is reached, in this case J. The second column then gives the sample size, i.e. for a single normal plan 80 on the same row as the J. If one continues along the row, the remainder of the sampling plan can be found by looking under the column with the specified acceptable quality level. Thus, for an acceptable quality level of 1%, we find an *Ac* (acceptance) number of 2 and a *Re* (rejection) number of 3. Similarly, for a double normal plan, the initial sample size is 50 with $Ac = 0$ and $Re = 3$. The second sample is also 50 with the revised values of $Ac = 3$ and $Re = 4$.

17.6.4.5 Selection of sample

It is important that the units selected for the sample are representative of the batch as a whole. Care must therefore be taken to select within the sample items taken from different areas of the batch.

17.6.4.6 Class of inspection

Normal inspection should initially be used, unless otherwise specified. This should continue to be used until:

(1) two out of five consecutive lots are rejected—then move to the tightened table until five consecutive lots are accepted, and then return to the normal table; *or*
(2) 10 consecutive lots have been accepted at the normal criteria—then move to the reduced table until one lot is rejected, and then return to the normal table.

17.6.4.7 Curtailment of sample testing

There will often be occasions before testing of the sample is complete on which the results being obtained make a final result *seem* probable. The desire to curtail testing at this point should be strongly resisted. Often full results are required to form a historical record of ongoing quality levels. Even where these are not being kept, the full result will tell clearly how well the normal quality level is being maintained.

It may appear permissible to stop testing once an accept/reject point is certain:

(1) if the reject number is reached before the full sample is tested; *or*
(2) if the number remaining is less than that required to reach a decision point.

If any such curtailment is to be allowed, it must take place under tight controls and be documented as such. It is often easier to insist on full testing than to administer this.

17.6.5 The operating-characteristic curve

Any sampling plan selected has a risk of producing a misleading result because it is based on only a portion of items. For each condition indicated by the sampling plan, there is a probability of either of two conditions being true, i.e. that indicated or the reverse.

These risks are important in two particular ways (see *Table 17.6*):

(1) a type I error can occur, i.e. the lot which is acceptable (at the acceptable quality level) is rejected by the producer;
(2) a type II error can occur, i.e. the customer receives a lot which contains more than the maximum defective level they are prepared to accept, i.e. the lot tolerant per cent defective (LTPD).

The risk over the complete range of possible defectives can be demonstrated using the operating-characteristic curve graph. One axis of the graph signifies the probability of any event occurring, the other the quality level, i.e. the percentage defective.

The curve represents the sampling plan. Any point on the graph shows the probability of the acceptance number (*Ac*) being the exact number of defectives found in the sample at that defective level.

If a sampling plan has no risk involved it will produce an operating-characteristic curve like the one shown in *Figure 17.26*. However, because we are dealing with samples, then

Table 17.6 Inspection transformation matrix

	True quality	
Condition indicated	*Good*	*Bad*
Good, batch accepted	OK	Error II
Bad, batch rejected	Error I	OK

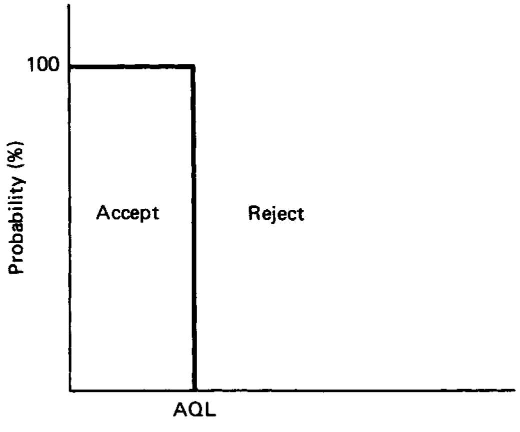


Figure 17.26 The ideal operating-characteristic curve. AQL, acceptable quality level

this ideal cannot be achieved, although we do wish the sampling plan to approach it as closely as possible.

17.6.5.1 Construction of the operating-characteristic curve

The starting point for drawing the operating-characteristic (OC) curve is the details of the sampling plan: the sample size n, and the acceptance number Ac. We then consult Poisson tables such as those given in Table A17.4 of the Appendix or Poisson charts such as the Thorndyke chart.

Before we can consult the Poisson table, we select the range of defectives that we wish to examine under the sampling plan. We then calculate the value of nd, which is simply the multiple of the sample size n and a set of defective levels d within the defective range. We then look up the Poisson table using the value of nd in the left-hand column against the sample value Ac, which heads each column. The resultant value is the cumulative probability P of finding Ac defectives or less in the sample at d defectives. It is this value that we plot to form the OC curve.

Example

Using the sampling plan for the batch of 1000, gave a sampling plan with $n = 80$ and $Ac = 2$.

d%	0.1	0.2	0.5	1.0	2.0	5.0	10.0
nd	0.08	0.16	0.4	0.8	1.6	4.0	8.0
P	1.00	0.99	0.99	0.95	0.78	0.24	0.01

It is obvious that the selected range of values for d could do with some extra values between 2 and 10% to be included to enable a smoother OC curve to be drawn, i.e.

d%	3.0	4.0	6.0	7.0	8.0	9.0
nd	2.4	3.2	4.8	5.6	6.4	7.2
P	0.57	0.38	0.14	0.08	0.05	0.02

The resultant value of the cumulative probability is plotted against the corresponding value of the defective to show the unique OC curve for this sampling plan (see *Figure 17.27*).

If we take the producer's risk as 5% and the customer's risk as 10%, then we can see that on this particular OC curve they have associated with them an acceptable quality level (AQL) of 1% and a lot tolerant per cent defective (LTPD) of 6.6%.

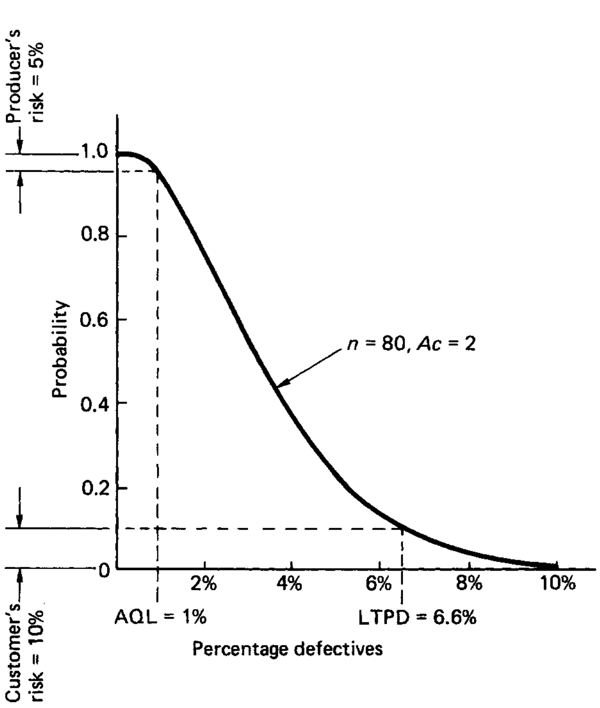


Figure 17.27 An operating characteristic curve. AQL, Acceptable quality level; LTPD, lot tolerant per cent defective

17.6.5.2 Effect of varying sampling-plan criteria on the operating-characteristic curve

Increasing the sample size n, even if keeping the acceptance number in the same proportion, will result in greater discrimination, but at an increased cost. The OC curve will tend towards the ideal rectangular shape (see *Figure 17.28*).

Reducing the acceptance number Ac, while holding the sample size at the same value, will mean more batches being rejected (see *Figure 17.29*). Not only will suspect batches be rejected, but so will many which are of satisfactory quality.

The effect of changing the inspection level is shown in *Figure 17.30*. Tightened inspection is more discriminatory, but costs more. Reduced inspection is the cheapest, but carries higher risks of passing lots with high defective rates.

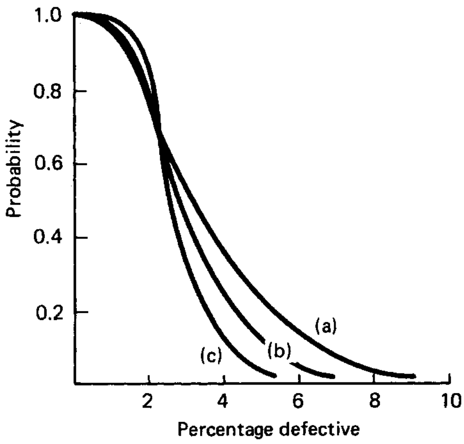


Figure 17.28 The effect on the operating-characteristic curve of increasing the sample size: (a) $n = 80$, $Ac = 2$; (b) $n = 160$, $Ac = 4$; (c) $n = 320$, $Ac = 8$

17.6.5.3 Outgoing quality level

If all the rejected batches are subjected to 100% inspection and the defects made good, then it is possible to calculate what the average outgoing quality level (AOQL) is over the range of the sampling plan. The AOQL is obtained by using the following formula, substituting in each case the appropriate values of the percentage defective d:

$$\text{AOQL} = \frac{P \times d}{P + [(1 - P) \times (1 - (d \div 100))]} \qquad (17.2)$$

Using the information from the sampling plan used above, we can substitute values in this formula. When there is a

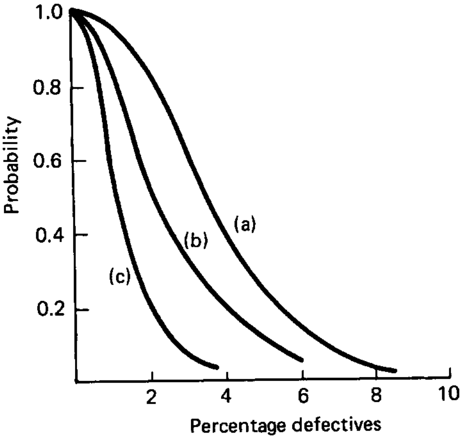


Figure 17.29 The effect on the operating-characteristic curve of varying the acceptance number: (a) n = 80, Ac = 2; (b) n = 80, Ac = 1; (c) n = 80, Ac = 0

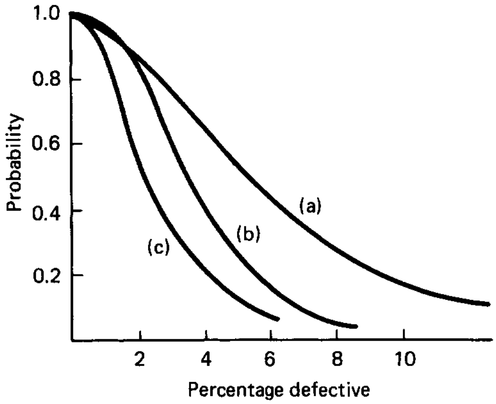


Figure 17.30 The effect on the operating-characteristic curve of varying the inspection level: (a) reduced inspection; (b) normal inspection; (c) tightened inspection

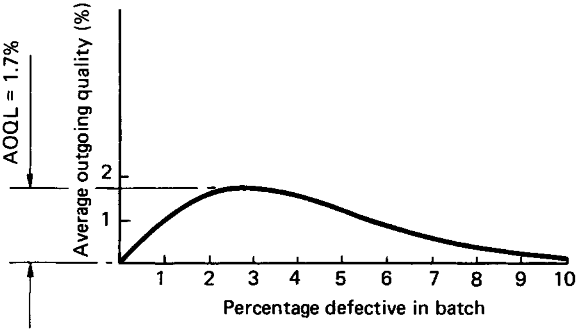


Figure 17.31 The outgoing quality level. AOQL, Average outgoing quality level

defective value of 1%, and an accumulative probability of 0.95, then:

$$\text{AOQL} = \frac{0.95 \times 1}{0.95 + [(1 - 0.95) \times (1 - (1 \div 100))]}$$
$$= 0.95$$

Covering the range of the sampling plan we can calculate the following relationships:

Lot defective	*AOQL*
0	0
0.5	0.50
1	0.95
2	1.56
3	1.73
4	1.56
5	1.25
6	0.89
7	0.60
8	0.43
9	0.20
10	0.11

These values are plotted in *Figure 17.31*. It can be seen that, if all rejected batches are subjected to 100% inspection, the maximum outgoing quality level occurs around the 3% incoming defective level, but is limited to a value much less than this, i.e. around 1.75% outgoing.

17.6.6 Calculating a non-standard quality plan

Whereas BS 6001: 1991 covers the normal requirements of acceptance sampling, there can be occasions when the customer requires a lot tolerant per cent defective (LTPD) different from that offered by the BS 6001: 1991 sampling plans. In these cases it is possible to work out a suitable sampling plan if the following factors are known:

(1) the desired producer's risk, α;
(2) the corresponding acceptable quality level, AQL;
(3) the desired customer's risk, β; and
(4) the corresponding LTPD.

We have to consult the Poisson probability charts (*Table A17.4* in the Appendix to this chapter) and tabulate the values of nd which lie under each of the acceptance numbers Ac for the desired risks α and β. We also produce a further column containing the ratio between these columns. When this ratio agrees with that between the AQL and the LTPD this gives us one part of the sampling plan—the acceptance number Ac. Knowing the value for nd we can then work back to the sample size as:

$$\frac{nd}{100} = \frac{\text{AQL} \times n}{100} = P \qquad (17.3)$$

Example

Using data from the operating-characteristic curve shown in *Figure 17.27*: α = 5%, AQL = 1%, β = 10%, LTPD = 6%. Values of nd at 95% and 10% and the ratio of these two nd values are given in *Table 17.7* against different Ac values.

The ratio LTPD/AQL = 6/1 = 6. The nearest value for this ratio in *Table 17.7* is that associated with an acceptance

Table 17.7 Data taken from *Figure 17.31* for use in the worked example of calculating a non-standard quality plan

Acceptance No., Ac	*Value of* nd *at* P = *95%*	*Value of* nd *at* P = *10%*	*Ratio of columns*
1	0.36	3.9	10.83
2	0.80	5.3	6.63
3	1.35	6.7	4.96
4	1.97	8.0	4.06
5	2.6	9.3	3.58

number of 2. Substituting the corresponding value of P, i.e. 0.8, in equation (17.3)

$n = (100 \times 0.8)/1$

$= 80$

Therefore, this method gives us a sampling plan with a sample size of 80 and an acceptance number of 2.

In this case, the calculated value is exactly the same as that suggested in BS 6001: 1991. This should not be surprising, as the data were derived from the operating-characteristic curve produced from the single sampling plan from BS 6001: 1991.

17.7 Statistical process control

In any production process, the important quality aspect is to ensure that the process continues within control. The use of statistical procedures and control charts enables this to be done in a cost-effective manner. The basis of the control lies in the continual calculation and charting of the actual results and comparing them with the probability of these results being within a normal frequency distribution.

17.7.1 The normal distribution

The normal distribution of an unbiased process should form a symmetrical, bell-shaped curve with tails extending, in theory, to infinity in both directions (see *Figure 17.16*). The important characteristics of the normal distribution are:

(1) the *mean*—this is simply the arithmetic average value of the variations; and
(2) the *distribution* of the values around the mean—this is denoted by the *standard deviation* of the distribution.

If we know the mean of a population μ, and its standard deviation σ, then we can easily determine the probability of any single value, or a range of values, being part of the same population by consulting statistical tables. *Figure 17.32* shows the relationship between multiples of the standard deviation and the proportion of its corresponding population being contained by it.

17.7.2 Variable control charts

Many processes produce items which are measurable in discrete quantities. In quality these measurements are termed 'variables' and they are particularly useful in controlling the variation of a process over time. It is always better to take more than one value when examining a process as the probability of an average value of a sample being different from the real population is less than that for one individual reading. This is due to the fact that the standard deviation σ_n of a set of samples is reduced from the standard deviation σ of a set of individual measurements in inverse proportion to the square root of the sample size n. This is shown in *Figure 17.33*, and is represented by the formula:

$$\sigma_n = \sigma/\sqrt{n} \qquad (17.4)$$

The technique for controlling variables involves taking small samples from the produce of a machine or process and plotting from these the following two control charts.

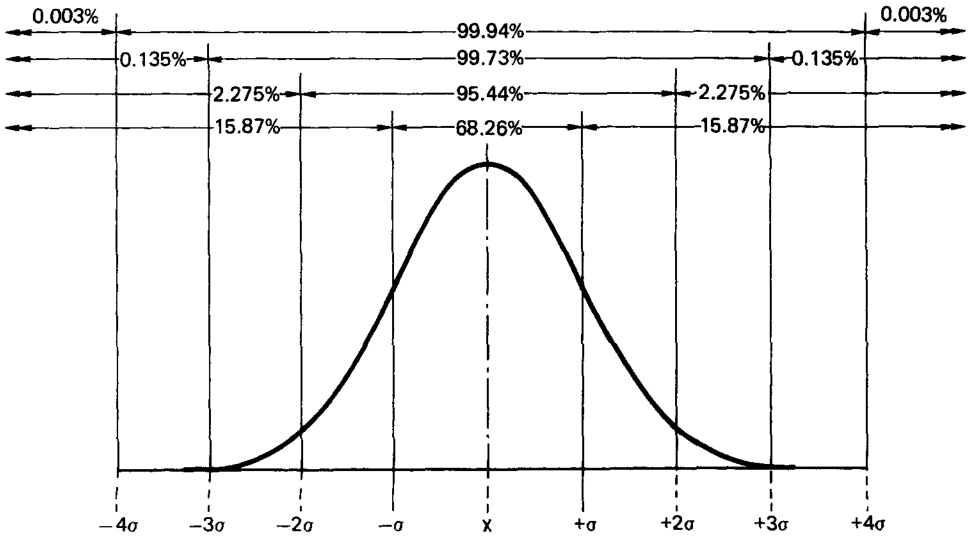


Figure 17.32 The coverage of a population by multiples of the standard deviation σ from the mean

18 Terotechnology and Maintenance

William T File

Contents

18.1 Management of assets 18/3
18.1.1 Assets in manufacturing 18/3
18.1.2 Terotechnology practice 18/3
18.1.3 Terotechnology and asset management 18/3
18.1.4 Maintenance and terotechnology 18/4
18.1.5 The asset register 18/4
18.1.6 Availability and related parameters 18/5

18.2 Life-cycle costing 18/7
18.2.1 Introduction to life-cycle costs 18/7
18.2.2 Life-cycle-cost process 18/7
18.2.3 Marginal analysis 18/8
18.2.4 Trade-off models 18/9
18.2.5 Minimising life-cycle costs 18/10
18.2.6 Practical difficulties and limitations of life-cycle costing 18/10

18.3 Plant selection and replacement 18/11
18.3.1 Plant-selection criteria 18/11
18.3.2 Plant procurement 18/12
18.3.3 After procurement of plant 18/12
18.3.4 Plant replacement 18/12

18.4 Measurement of the effectiveness of maintenance 18/13
18.4.1 The need for measuring maintenance 18/13
18.4.2 Maintenance costs 18/13
18.4.3 Maintenance indices 18/14

18.5 Operational aspects of maintenance 18/14
18.5.1 Maintenance planning and control 18/14
18.5.2 Computers and maintenance 18/15
18.5.3 The profit-centre concept 18/15
18.5.4 Direct and indirect costs of maintenance 18/16
18.5.5 Maintenance revenue 18/16
18.5.6 Pricing of maintenance 18/17
18.5.7 Maintenance contracts 18/18
18.5.8 Modifications 18/18
18.5.9 Refurbishing of parts and equipment 18/18

18.6 Preventive maintenance 18/19
18.6.1 Classifying types of maintenance 18/19
18.6.2 Failure patterns 18/20
18.6.3 Establishing preventive-maintenance routines 18/20

18.7 Condition monitoring 18/21
18.7.1 Introduction to condition monitoring 18/21
18.7.2 Alternative condition-monitoring strategies 18/22
18.7.3 Condition-monitoring applications and sensing equipment 18/22

18.8 Inventory and maintenance 18/24
18.8.1 Inventory objectives versus maintenance ideals 18/24
18.8.2 Cost of stock holding 18/24
18.8.3 Safety stock and reorder quantities 18/25

18.1 Management of assets

18.1.1 Assets in manufacturing

In manufacturing terms, 'assets' may be hand tools, machine tools, process plant, manufacturing premises, warehousing, raw materials, work in progress and finished goods awaiting despatch. The definition may also be extended to human resources, the effective utilisation of which are usually manufacturing's principal asset.

Effective utilisation of human resources depends on the availability of the 'tools' of manufacturing, and the ultimate objective must be to have all the manufacturing equipment available for use 100% of the time. In reality, this figure is never likely to be achieved, but figures approaching 100% should be the target. Obviously, in common with all other hardware, 'you get what you pay for!' and availability is, in reality, a trade off with cost. This is discussed further later in this chapter.

The hardware elements of the list above (as opposed to the human resources) represent the capital employed in the production of goods. Before being put into use, these same hardware elements must be:

(1) specified,
(2) selected,
(3) designed,
(4) built,
(5) purchased,
(6) installed, and
(7) commissioned.

At some time in the future these hardware elements will have outlived their economic life and will either be sold or scrapped. This is the life cycle of manufacturing equipment; it begins when its acquisition is first considered and ends when disposal is finally completed. Typical relative costs associatd with each phase are shown diagrammatically in *Figure 18.1*.

It should be noted that the digram assumes some residual value in the equipment at the end of its useful life, and its sale potential will lower the cumulative life cost on disposal. One thing that is very clear from the diagram is that the costs in the early phases are relatively low compared with building/purchasing and operational/maintenance costs, so time spent ensuring that equipment is correctly specified will clearly be advantageous in the longer term.

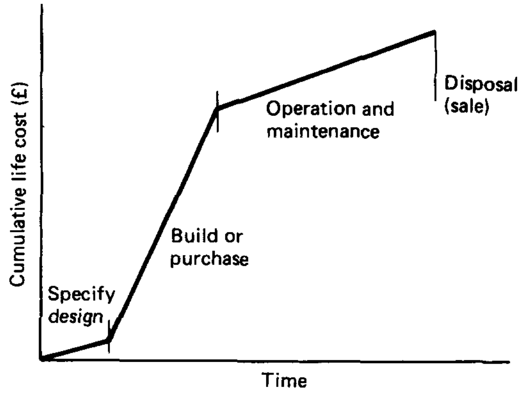


Figure 18.1 Cumulative life-cycle cost of an asset

18.1.2 Terotechnology practice

The term 'terotechnology' was initiated in the 1970s as a result of a project commissioned by the British Government to study and report on the state of maintenance in industry. As a result of the initial report of this study group, its terms of reference were widened. It was considered that maintenance could not be considered in isolation from other elements of engineering and the costs involved in supporting a manufacturing process. More recently, a British Standard, BS 3843: 1991 *Guide to Terotechnology (The Economic Management of Assets)* has been published. Part 1 introduces the concept, Part 2 describes the techniques involved and Part 3 continues with short notes on the use and application of the various techniques.

Terotechnology is a combination of management, financial, engineering, building and other practices applied to physical assets in pursuit of economic life costs. In practice, it is concerned with the specification and design for reliability and maintainability of plant, machinery, equipment, buildings and structures, with their installation, commissioning, maintenance, modification and replacement, and with feedback of information on design, performance and costs.

The key words in the above definition are 'economic life cycle costs'. 'Economic' implies that it may not necessarily be the cheapest, as in all good designs there is a specification to be met and inevitably a trade-off situation arises where money must be paid for quality or reliability.

The successful application of terotechnology principles to the management of assets involves all relevant functions within a company and depends on the systems in place to implement it. Terotechnology also requires the commitment of senior management to the process, as well as the commitment to feed back cost performance to enable comparison with the original targets set at the time of the initial specification and appraisal. In turn, this information will be utilised as a reference database for the purchasing or manufacture of future assets.

18.1.3 Terotechnology and asset management

Organisations within companies differ for many reasons, e.g. the size of the company, the diversity of products or services, and the geographical location of premises. Systems generally tend to be linked to organisation structures, and terotechnology, being a management system, must take due account of the organisation imposed by management. Having said that, such a system must be designed for those who will use it.

Companies which are organised to control their output by some form of product management system, whether through line or matrix management, have the simplest task, but for others there is no need to consider any reorganisation, providing line management and others accept the principles of terotechnology. Small companies will find that an informal system linked through their normal management processes will be sufficient. Larger companies, particularly if several different sites are involved with the same product or service, may perhaps require a nucleus of dedicated staff. However, it is essential that the costs of any asset-management system be kept to a minimum and that data presented to management are adequate to make the necessary decisions, but not excessive.

The majority of data required for the planning associated with terotechnology already exists in most companies. It will simply require collating and formatting to present a coherent life-cycle analysis.

The activities necessary for the implementation of terotechnology principles include the following.

(1) Decide which of the economically viable projects being considered should go forward.
(2) Carry out a detailed investment appraisal for these projects and then set targets for production costs and income from each.
(3) Decide what physical assets are needed and how to use them to achieve the targets over a specified period, taking into account their forecast costs of acquisition, production, maintenance and disposal.
(4) Specify the asset to be acquired and its performance.
(5) Assess tenders for the asset in terms of its life-cycle costs and economic performance. Select the tender which is preferable from a terotechnological point of view.
(6) Acquire, install and commission these assets.
(7) Operate and maintain the assets in a way which gives optimum economical performance.
(8) Monitor the use of the assets in terms of performance and cost.
(9) Modify to improve the economical performance of the assets using information from monitoring.
(10) Use the data collected to contribute to a decision regarding any replacements needed and the optimum time for these replacements to be carried out.

18.1.4 Maintenance and terotechnology

The application of terotechnology principles requires expertise in a variety of different job functions, and it can be seen from *Figure 18.2* that maintenance is the core of this process, in as much as the total maintenance function is included in terotechnology. By contrast, for example, the design function would be responsible for the technical decisions on alternative ways of meeting a product or process specification, independent of terotechnology, except where the cost implications of the decision may impact total life costs.

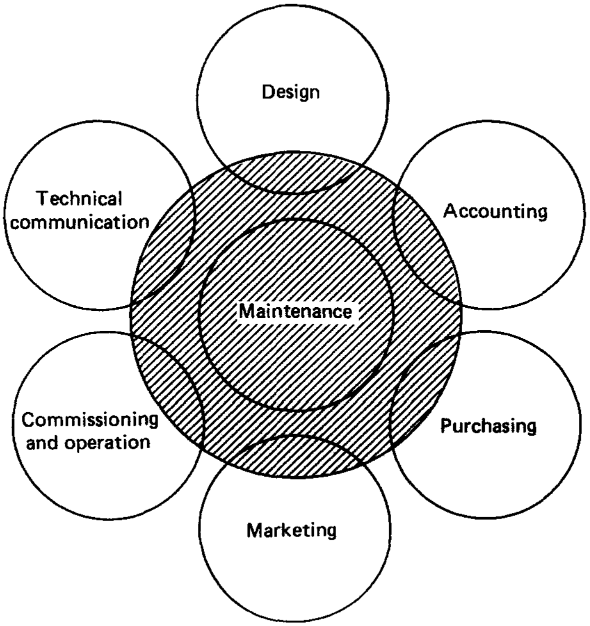


Figure 18.2 Relationship between terotechnology (hatched area) and different job functions

18.1.4.1 *The role of the maintenance manager*

The maintenance manager is, by definition, responsible for the care and satisfactory performance of the company's assets. He/she therefore plays a key role in controlling asset life costs and his/her involvement must begin at the tender and acquisition phase for new plant or process machinery. The manager has a valuable store of knowledge of the company's existing resources and capabilities, which, if tapped early enough, can be used to its best advantage in minimising life costs.

The following guidelines are among those which are key to the successful application of terotechnology as it affects the maintenance manager.

(1) The contribution from maintenance begins before plant is purchased.
(2) Information from the maintenance function should be used at the investment appraisal and tender assessment stages in the acquisition cycle.
(3) Adequate technical data and maintenance recommendations in an acceptable form should always be provided by the supplier of a new asset.
(4) Modern management and maintenance techniques should be applied to the company's assets to provide the highest availability consistent with overall economic performance over the whole life cycle.
(5) Comprehensive maintenance performance and cost information should be gathered and used to seek means of improvement.
(6) The maintenance function should make information concerning performance and costs associated with assets available to other functions within the company.
(7) Maintenance personnel should be encouraged to be aware of the cost implications of their actions.
(8) The maintenance manager should learn the 'language' of accountants and general management.

18.1.5 The asset register

The majority of organisations keep some records of the equipment that they use, either in hard-copy form or on a computer file. These can take various forms, but often only contain the information available at the time of purchase. Much valuable data is lost if these records are not updated regularly. This is usually due to the fact that the department holding the records has no direct interest in performance during the life of the equipment.

The maintenance manager, by virtue of his role in keeping his company's assets in a good state of repair, is ideally placed to be custodian of the company equipment register. On this register can be entered a summary of the maintenance cost and performance data relevant to the equipment concerned. This information will provide the basic input for life-cycle costing when the time comes to consider replacing the equipment. The information normally kept on the equipment register will not be in sufficient detail for day-to-day maintenance planning. However, there is an obvious advantage in being able periodically (and automatically) to transfer the total maintenance cost data from the maintenance reporting system to the equipment register, thus ensuring that it is always up to date and compatible with the annual operating plan. The register will be in the economics of each individual year, but with care it can provide an ideal base for long-term planning. *Figure 18.3* shows a typical format for an asset register.

Equipment description								
Year	*Planned costs*	*Actual costs*			*Performance*			*Comments*
		Labour	*Spares*	*Total / Planned*	*Planned*	*Actual*	*Actual / Planned*	
Initial cost →		—	—	—	—	—	—	
1								
2								
3								

Figure 18.3 Example of an asset register

18.1.6 Availability and related parameters

A recent Department of Trade and Industry survey showed that £8 billion is spent every year on maintaining direct production assets in the UK. This figure must be a fraction of the costs lost whilst production is interrupted for the maintenance process. In almost any situation, therefore, this represents a serious drain on the financial resources of the company and its ability to meet production targets. Few of the companies in the study attempted to quantify the indirect costs of poor and inefficient maintenance. Much of the indirect cost is incurred by the downtime of the equipment waiting for the repair to be completed, parts to be supplied, or perhaps just waiting for the maintenance engineer himself. Consequently, inadequate maintenance in one form or another results in poor manufacturing performance and, ultimately, in dissatisfied customers and lost markets.

It has been said that the competitive advantage gained by a company from expensive and complex manufacturing processes is only as good as the company's ability to ensure optimum availability. Whilst 100% availability is often desirable, it is seldom achievable. Perhaps the best example of 100% availability is to be seen in modern aircraft, where complete back-up systems or stand-by systems are available to ensure the safe arrival of passengers at their destination. From the manufacturing point of view, such back-up and stand-by systems would probably not make economic sense and, therefore, we are looking at something less than 100% availability. How much less is a question of economic and financial trade-offs.

Availability, therefore, is perhaps the most user oriented of all the indices, being in effect the user's perception of the reliability of the equipment.

$$\text{Availability} = \frac{\text{Possible operating time} - \text{Down time}}{\text{Possible operating time}}$$

In order to understand the concept of availability, it is important to understand the allocation of time (see *Figure 18.4*). One significant point which can be seen from *Figure 18.4* is the fact that maintenance can be performed either during the working day or during non-operative time. In the latter case there is the advantage that maintenance does not affect availability and, therefore, provides an interesting opportunity for restricting downtime by preventive maintenance done outside working hours. When purchasing new equipment, care should be taken to ensure that the derivation of the figures for the 'possible operating time' are those which correspond to the user demand on the equipment.

The costs associated with availability are shown in *Figure 18.5*. As can be seen from the figure, there is an optimum economic point for availability. As the availability approaches 100% the downtime costs will decrease and, inevitably, the maintenance costs will increase. The overall picture is, however, to some extent dependent on the capital cost of the equipment in the first place and, as can be appreciated, paying additional costs up front for easier maintenance and improved reliability would significantly alter the percentage points on the availability scale in *Figure 18.5*.

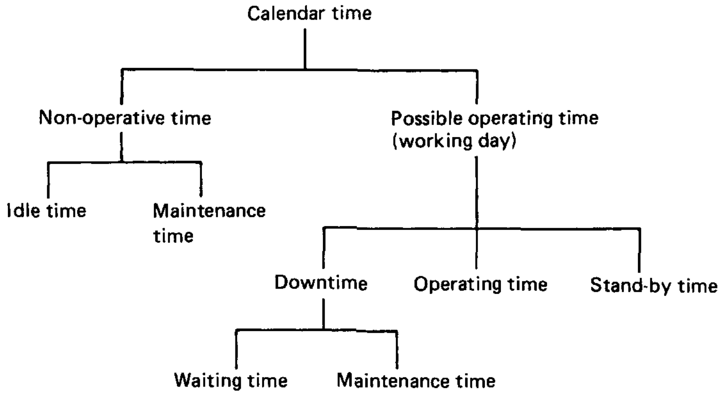


Figure 18.4 Time elements relating to maintenance

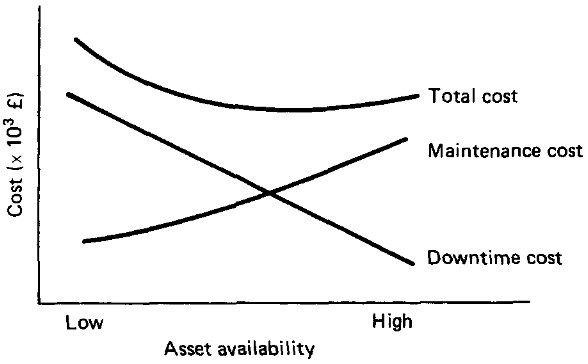


Figure 18.5 Costs associated with availability

Let us therefore look more closely at downtime.

Downtime = No. of failures × Time taken to restore equipment to its operational state

Obviously, downtime can be reduced by improving one or other of the factors on the right-hand side of the equation. The number of failures is directly dependent on the reliability specification of the equipment, whilst the time taken to restore the equipment will depend on:

(1) waiting time, and
(2) maintenance time.

In turn, the maintenance time will depend on:

(1) the skill of the maintenance engineer,
(2) adequate tools and spares, and
(3) the ease with which the equipment can be maintained.

This last factor is called *maintainability* and is defined as the property of the design which determines the effectiveness by which an equipment can be retained or restored to a state of readiness for operational use.

18.1.6.1 Reliability

Reliability can be expressed as the number of failures in a time period, which may be the expected life of the equipment. More often it is expressed as a failure rate, i.e. the *mean time between failures* (MTBF). This has the advantage that if, as usually happens, the failure rate varies over the life of the equipment, then different rates can be included in the specification related to the age of the equipment. These are usually designated as the *initial failure rate* (immediately after installation) and the *mature failure rate* (at some later defined time in the equipment life) (see *Figure 18.6*).

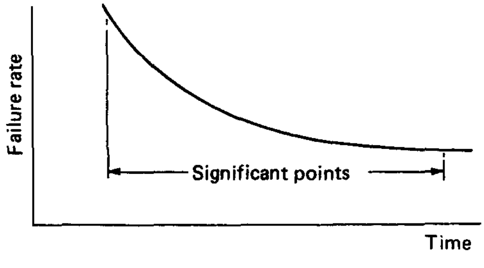


Figure 18.6 Specification of failure rates

18.1.6.2 The maximum and mean time to repair

These are two other important maintenance parameters. These will significantly influence the maintainability features of the equipment and, if coupled with the waiting time, can determine the level of spares inventory to be held, as well as geographical location relative to the equipment (see Section 18.8).

18.1.6.3 Operating cycle

Availability, reliability (and hence maintenance costs) and also the life of the equipment can show significant change in values with different operating and stand-by relationships. It is therefore essential that a clear understanding of the intended use of the equipment is established at an early stage in specification development. This can be demonstrated using a vending machine as an example.

Suppose the vending machine manufacturer quotes an availability of 98% based on reliability data of one equipment failure per 10 000 cups of beverages dispensed, how many cups per day does this relate to and is the day 8 or 24 h long? If it is a normal 8 h working day, then this availability figure would give a downtime of 6.4 h in 40 days, say one breakdown in 40 days allowing 6.4 h to cover both the response time and work time of the engineer.

Let us also assume that the vending machine is installed in a factory where there is shift working and a further 500 cups of beverage are dispensed during the two night shifts. The quoted availability will be the same at 98%, but a breakdown will occur every 13.3 days. Unfortunately (for the night shift), the maintenance engineer for the vending machine is unlikely to be on stand-by overnight. The probability of a failure during this time is 0.67, with a resulting penalty on response time of an average of at least 8 h. This has a considerable effect on the availability:

$$\text{Availability} = \frac{(13.3 \times 24) - [0.67(6.4 + 8) + (0.33 \times 6.4)]}{13.3 \times 24}$$

$$= 0.963 \text{ or } 96.3\%$$

Note that if the vending machine was operational over the weekends and the engineer was not available until Monday, then the availability would have been even lower.

We have already seen that the failure rate (or MTBF) makes a significant contribution to the availability of a piece of equipment, and in the examples above it was assumed that it was constant at 1 per 10 000 cups. In practice, this would have been correct in this instance because the average number of cups of beverage dispensed during each shift was also constant. However, had the machine been left in the 'ready' state all night, but principally only used during the 8 h day shift, its usage characteristics would change. In all probability the failure rate would have worsened marginally against the previous figure of 1 per 40 days due to the prolonged period over which the water was being kept hot in readiness for instant dispensing.

18.1.6.4 Usage rate

As can be seen from the above, usage rate is an important factor related to failure rates and *Figure 18.7* demonstrates the typical variation of MTBF with usage rate. The relationship shown in *Figure 18.7* is applicable to most situations, given that average (not peak or total) usage rate data are used. If the design is optimised for the intended usage, then the normal operating range would be expected to be found in the flat part of the curve.

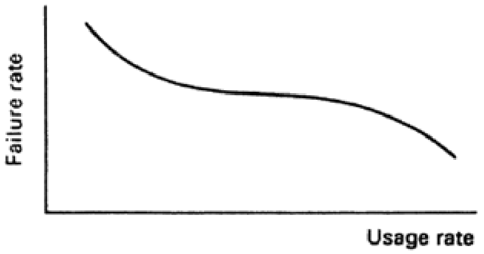


Figure 18.7 Typical variation in failure rate with usage rate

18.2 Life-cycle costing

18.2.1 Introduction to life-cycle costs

It is often assumed that the cost of maintaining equipment is constant over its expected life. This assumption has the drawback that it presupposes that the cost factors associated with maintenance remain unchanged over the years, and does not take infant mortality or wear out into account. Whilst performance over life may also vary according to a predefined pattern, various other factors relating to the cost of money do not vary in the same way, such as wage rates, inflation, etc. When a detailed justification for purchase of a new equipment is being prepared, then these factors must be considered. The process of life-cycle costing provides this opportunity, taking into account not only maintenance costs, but also all the other costs associated with the purchase and operation of an equipment that may be necessary.

If the equipment is to be purchased, for example, there will probably be capital costs in the first year and quite possibly a significant loan on which to pay interest and capital in subsequent years. Then there are inflation effects to consider, not only on the load, but also on the costs of maintenance itself. So, what does the equipment really cost, and what impact will its operational costs have on the company profits over a number of years?

Life-cycle costs The life-cycle costs associated with a piece of equipment are the costs of acquiring, using, keeping it in operational condition and final disposal (see *Figure 18.8*). These costs include the costs of feasibility studies, research, development, design, manufacture, installation, commissioning, operation, maintenance and replacement, and any credit for disposal. The costs of any supporting actions generated by these activities would also be included, such as training, documentation, tools and test equipment, inventory holdings,

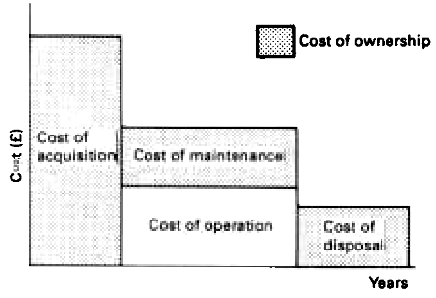


Figure 18.8 Histogram of life-cycle costs

and any other additional facilities required as a direct result of the use of a piece of equipment. It can be stated that:

Life-cycle cost = Cost of operation + Cost of ownership

Cost of operation This comprises the direct operational costs such as energy, materials and labour used in the production of goods or services.

Cost of ownership This includes all the items defined above for life-cycle costs, with the exception of those listed under cost of operation. Cost of ownership can be subdivided into three parts, which are self-explicit: *cost of acquisition, cost of maintenance* and *cost of disposal.* The cost of ownership is, therefore, most relevant to the maintenance function.

Conventional company accounting is structured in such a way that direct costs are divided between those related to a product or service and indirect costs are shared as an overhead over the entire range of products or services offered. Life-cycle costing, which is a management accounting process, is intended to present the costs so that they are of direct assistance in the making of management decisions related to a particular product or service. Whilst the two approaches are entirely separate, they must be consistent in their treatment of cost data in order to be acceptable to all functions. In general terms this includes:

(1) labour overheads and utilisation factors;
(2) differing labour rates for different skill levels;
(3) parts overheads for inventory and distribution costs; and
(4) consistency with procedures for compiling operating budgets and long-term plans.

The principal use of the life-cycle-cost process is for comparative studies for alternative procedures, equipment purchases, trade-off decisions or similar exercises where a decision must be made on future courses of action. Situations may arise where trade-offs may be necessary between cost of ownership and cost of operation, or perhaps cost elements within either of these categories. A simple example of this would be the relatively high cost of installing thermal insulation, which would be offset by lower energy costs during subsequent years.

18.2.2 Life-cycle-cost process

Life-cycle costing requires clear terms of reference before embarking on any gathering of data towards decision-making. The results will only be as good as the input, and it is essential to obtain the commitment of all relevant functions involved to ensure the validity of the data obtained. Life-cycle costing is essentially a multifunctional task but, because of the central role of the maintenance manager in his/her responsibility for the upkeep of assets, he/she can well be the coordinating function in such a decision-making process.

Life-cycle costing requires:

(1) clear objectives;
(2) recognition of all alternative courses of action;
(3) realistic assumptions and forecasting of all data elements;
(4) consistent approach to the use and application of all data;
(5) analysis of difference in cost of each course of action, and all possible outcomes; and
(6) comparison of the final cost data with the objectives.

These requirements cannot be overemphasised, as life-cycle costing can be a time-consuming exercise and, if not approached correctly and comprehensively, will produce mis-

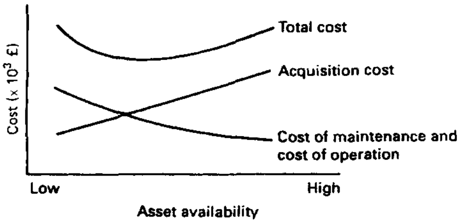


Figure 18.9 Optimum life-cycle costs

leading results. When reviewing alternative courses of action, care should be taken to ensure that all assumptions made are varied according to the appropriate course of action and, where these are difficult to forecast, best and worst situations should be evaluated. Forecasting techniques used should be those which are already the standard within the company, and accepted by all other functions affected by the outcome of the investment decision-making process of which life-cycle costing will form a part.

A typical relationship showing the effect of asset availability with different acquisition costs and costs of maintenance and operation is shown in *Figure 18.9* (note the similarity with *Figure 18.5*). This demonstrates the point that within reason it is often worth incurring higher initial costs with a view to minimising the ongoing maintenance and operating costs.

Speculation phase When first considering the purchase of new equipment a graph such as that shown in *Figure 18.9* is a good way to demonstrate the cost impact of various alternative options. As these are reviewed, they may be plotted on this basis to determine the optimum cost performance of the equipment. If a number of alternatives is available which have projected life costs falling within the central low area of the total cost curve, then a considerable improvement in availability can be achieved by selecting one with the higher acquisition cost, without significantly affecting the total life cost.

Appraisal phase Having narrowed the field to perhaps two or three alternatives, detailed cash-flow projections should be prepared on a year-by-year basis, introducing alternative maintenance policies, and discounted to net present value (NPV). The discount factor used would normally be provided by the accounts department and would depend on a combination of inflation and market interest rate projections.

To show the effect of a discount factor, consider investing £100 today and assume that the best interest rate obtained will be 10%. This will provide £110 in 1 years time and £121 in 2 years time. It is said then, that the NPV of £121 in 2 years time is £100, given an interest rate of 10%. In other words, £100 is needed today to achieve £121 in 2 years time, the discount rate for the 2 years is then 100/121 = 0.83.

This can best be demonstrated by an example. Suppose a vehicle is due for replacement and a number of alternative makes with different specifications is available. The vehicle chosen after reviewing the market is a type costing £30 000, and this is entered in the first column in *Table 18.1* at 'year 0'. Had there been a loan involved then this would have been extended to perhaps 3 years together with appropriate interest payments. The operating costs are constant for the period under review at a total of £4000 per annum. Maintenance costs are shown to be increasing gradually from 'year 3' onwards, and coincident with this there are revenue losses shown due to the vehicle being off the road whilst under repair. The first row in the table is at constant economics at 'year 0', which is effectively today. For each successive year completed a 10% discount is estimated. This provides a cumulative total at NPV of £77 300, as the life cost taken over a 6-year span.

As any such project proceeds, the original factors assumed should be monitored and updated. They can then be used as a basis for monitoring and controlling performance and costs. Annual budgets, operating and long-term plans will often disguise fundamental project-related problems, including those of maintenance costs, which can only be highlighted by a life-cycle-cost review.

18.2.3 Marginal analysis

Supposing that there will be an option to overhaul the vehicle (used in the example in *Table 18.1*) in year 4 at a cost of £8000, with maintenance costs and therefore revenue losses being reduced thereafter, then a new life cost could be established. This time, however, any costs which are common to both situations, such as the original capital costs and the operating costs, will be ignored and the calculation based on the variable costs. This is known as 'marginal costing' and is usually used on a comparative basis as shown in *Table 18.2*. If the cumulative total shows a gain, then it should be a worthwhile proposition, assuming that the underlying assumptions are sound.

It is often worth varying the key assumptions to ensure that they are not critical to achieving the calculated result. For example, supposing that the maintenance and revenue cost savings projected above are not realised, and they are 30% higher than expected. Then substituting the new data in *Table 18.2* would show an expected cumulative loss at the end of year 6 of £100 instead of the expected £3600 gain. If there is a

Table 18.1 Example of life-cycle costing

Year	*Capital* ($\times 10^3$ £)	*Operating costs* ($\times 10^3$ £)	*Maintenance costs* ($\times 10^3$ £)	*Revenue losses* ($\times 10^3$ £)	*Total (constant)* ($\times 10^3$ £)	*Discount*	*Total (NPV)* ($\times 10^3$ £)
0	30				30.0		30.0
1		4.0	3.0		7.0	0.91	6.4
2		4.0	3.0		7.0	0.83	5.8
3		4.0	3.6	2.0	9.6	0.75	7.2
4		4.0	4.3	4.0	12.3	0.68	8.4
5		4.0	5.2	6.0	15.2	0.62	9.4
6		4.0	6.2	8.0	18.2	0.56	10.2
Cumulative total							77.3

Table 18.2 Marginal cost—overhaul option

Year	Base case		Option		Gain	Discount	Gain at NPV
	Maintenance cost ($\times 10^3$ £)	Revenue losses ($\times 10^3$ £)	Maintenance cost ($\times 10^3$ £)	Revenue losses ($\times 10^3$ £)	($\times 10^3$ £)		($\times 10^3$ £)
0					0		0
1	3.0		3.0		0	0.91	0
2	3.0		3.0		0	0.83	0
3	3.6	2.0	3.6	2.0	0	0.75	0
4	4.3	4.0	11.0	2.0	4.7*	0.68	3.2*
5	5.2	6.0	3.6	2.0	5.6	0.62	3.5
6	6.2	8.0	4.3	4.0	5.9	0.56	3.3
Cumulative total							3.6

* Loss.

high probability of this occurring, then obviously it would be unwise to make the investment in overhauling the vehicle in year 4.

18.2.4 Trade-off models

In *Table 18.2* and the accompanying narrative a trade-off situation was studied using the marginal costing technique, which compared two situations, one the 'base case' and the second introducing a major overhaul in year 4 of operation. If a number of such alternative strategies were to be considered, it would be worthwhile considering creating a trade-off model which could be used by flexing the base case. In the example quoted in *Table 18.3* the costings used in *Table 18.2* are reproduced individually at NPV, rather than adding them at yearly values and adjusting the totals by the discount factor.

The effect is to put all the costs at todays value of money and, therefore, direct comparisons can be made between the various cost factors, irrespective of when they occur. Hence the maintenance costs of £17 400 over the 6 year vehicle life is now directly comparable with the cost of the vehicle at £30 000 and it can be said, for example, that an additional £5000 can be spent on the vehicle if a saving of more than the same amount can be made on operating costs, maintenance costs or reduced revenue losses. This is, however, an investment decision, and cash-flow considerations at the time of purchase might prohibit this action because the return may only be realised later during the 6-year life of the vehicle.

As a further example, suppose that the opportunity exists to reduce the maintenance costs for a range of pumps by

Table18.3 Costing of individual items at NPV

Year	Capital ($\times 10^3$ £)	Operating costs ($\times 10^3$ £)	Maintenance costs ($\times 10^3$ £)	Revenue losses ($\times 10^3$ £)	Total (NPV) ($\times 10^3$ £)
0	30				30.0
1		3.7	2.7		6.4
2		3.3	2.5		5.8
3		3.0	2.7	1.5	7.2
4		2.7	2.9	2.7	8.3
5		2.5	3.2	3.7	9.4
6		2.3	3.4	4.5	10.2
Total	30	17.5	17.4	12.4	77.3

additional capital investment to improve maintainability. It is assumed that the business under consideration is a major user of these pumps and, therefore, both the capital investment for the equipment and the maintenance costs are incurred by the same company.

Capital cost of each pump	£400
Installation rate of pump	600 per year for 3 years
Maintenance cost per pump year	£70
Total pump project cost (see *Table 18.4*)	= 597 000 + 314 000 = £911 000

In this particular instance, total life-cost figures, whilst understandable, are not of any immediate practical use. It is only a straightforward mathematical step to determine the direct relationship between, say, maintenance cost per pump year and capital cost per pump.

£1000 change in capital expenditure $= \dfrac{10^3}{1800}$

$= £0.55$ per pump

£1000 change in maintenance costs $= \dfrac{10^3}{6300}$

$= £0.159$ per pump year

At NPV,

Average capital cost per pump $= \dfrac{597 \times 10^3}{1800}$

$= £331.7$

Average maintenance cost per pump year $= \dfrac{314 \times 10^3}{6300}$

$= £49.8$

At todays value of money, therefore,

£1000 change in capital expenditure $= 0.55 \times \dfrac{400}{331.7}$

$= £0.66$ per pump

£1000 change in maintenance costs $= 0.159 \times \dfrac{70}{49.8}$

$= £0.22$ per pump year

Table 18.4 Pump project costing

Year	*Population*	*Pump-years*	*Capital cost* ($\times 10^3$ £)	*Maintenance cost* ($\times 10^3$ £)	*Discount factor*	*Capital NPV* ($\times 10^3$ £)	*Maintenance NPV* ($\times 10^3$ £)
0	0						
1	600	300	240	21	0.91	218	19.5
2	1200	900	240	63	0.83	199	52.3
3	1800	1500	240	105	0.75	180	78.8
4	1800	1800	—	126	0.68	—	85.7
5	1800	1800	—	126	0.62	—	78.1
Total		6300				597	314

Therefore, the total expenditure over 5 years will be reduced if less than £3 (i.e. £0.66/0.22) is added to the capital cost of each pump to save £1 per pump-year in maintenance cost, at todays value of money.

18.2.5 Minimising life-cycle costs

When considering the alternative options open to management regarding any proposed project, the principal objective will be to minimise life-cycle costs without degrading the required performance. It may well be that a lower performance is acceptable if the life-cost savings are significantly high, but this is a trade-off that can only be made after reviewing the individual circumstances.

Whether it is for this reason, or any other, the maintenance manager will probably be faced with the problem of optimising his/her approach rather than minimising it. Modification or overhaul of existing equipment is, therefore, an optimum measure which economically may make good sense when capital expenditure on new equipment either cannot be justified or cannot be afforded. In *Figure 18.10* the annual maintenance costs of a typical piece of equipment are plotted cumulatively in constant economics, showing an increasingly steep curve as time progresses.

The introduction of a modification, shown at point A in *Figure 18.10*, will improve reliability and, therefore, tend to reduce the gradient of the ongoing cumulative maintenance cost. It is important that such a modification is introduced sufficiently early in the life of the equipment to ensure that the cost of it (represented by the vertical rise in cumulative cost) is

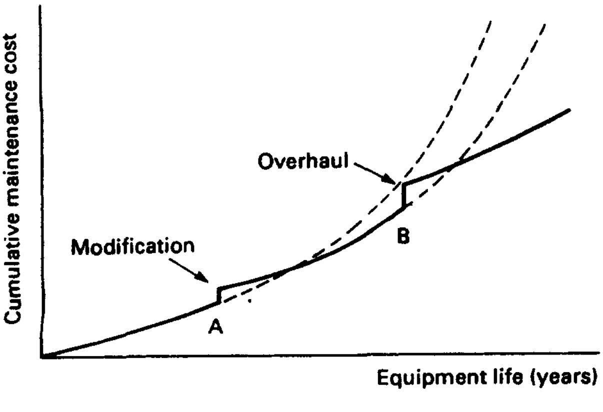


Figure 18.10 Plot of maintenance costs versus time

recovered by the subsequent maintenance savings. If there is any doubt, then a marginal analysis should be undertaken to satisfy management of the wisdom of their actions.

The steep rise as wear out of the equipment begins to show on the graph can be checked (or at best, delayed) by overhauling the equipment, and the effect of this is shown at point B in *Figure 18.10*.

18.2.6 Practical difficulties and limitations of life-cycle costing

As has been explained earlier, life-cycle costing is a project-related process which crosses all departmental boundaries within an organisation. One of the major difficulties which this creates, therefore, is to ensure that all data collected are complete and valid. It is very easy to overlook some of the less obvious implications of the project under investigation, particularly when considering a situation some years ahead.

When reviewing alternative courses of action, it is essential that all those providing data understand the assumptions being made and that their input is compatible with the course of action proposed. If these assumptions are difficult to forecast accurately, then a range should be evaluated.

The accuracy of the data projected for future years can also significantly affect the outcome of a life-cycle-cost exercise, and in this respect it is essential to adhere to existing forecasting techniques which are already a company standard and accepted by other functions.

Any decision which may be taken based on a marginal result must also be tested by making a small percentage variation to key input factors, to determine their sensitivity to the assumptions being made.

The company accountant's view may also constrain the introduction of a life-cycle-cost study, because his/her main concern is the annual budget and its impact on the profit-and-loss account. It is typical in many maintenance related projects to invest heavily in early years to gain the financial benefits later on. However, a significant investment can often give rise to a cash-flow problem.

For completeness, it should be mentioned here that there are three other methods of evaluating investments in projects which are in common use.

Return on investment (ROI) This is a simple calculation, expressing as a percentage, the savings or incremental earnings of an investment, divided by its cost. It has the disadvantage that it ignores both the effect of time and cash flows.

Payback This is an improvement over ROI in that it allows for the time element by calculating the length of time it takes

Table 18.5 Table of comparative accounting methods (all values in $\times10^3$ £)

	Year			*Total*
	1	*2*	*3*	
Gain/loss, current accounting	(4.7)	5.6	5.9	6.8
Discount (10%)	0.91	0.83	0.75	—
LCC	(4.3)	4.6	4.4	4.7
ROI of 8.0	3.3	5.6	5.9	85% (14.8)
Payback on 8.0	3.3	8.9	—	>2 years
IRR	(2.5)	1.6	0.9	85% (0)

LCC, life-cycle-cost; ROI, return on investment; IRR, internal rate of return.

to recover the initial investment cost of a project. It is normally expressed in constant economics.

Internal rate of return (IRR) The IRR factor is often used as a standard within a company to assess the financial benefits of new investments, but the establishment of this factor together with the calculations needed to establish it are best left to the accountants! It is a development of the life-cycle-costing process, and the IRR is defined as the percentage discount rate which must be achieved to give a zero net present value over a specified time. (See the life-cycle-costing example associated with *Table 18.1*, in order to understand this terminology.)

To understand the differences in presentation by using the above accounting methods, calculations for each method over a 3-year period are given in *Table 18.5*. The figures used are those of the overhaul option considered in *Table 18.2*, starting at the commencement of year 4 as 'today'.

Finally, if a decision is accepted, based on a life-cycle-cost analysis, then it is essential to follow it through with periodic reviews. As the project proceeds, all the relevant input and output factors should be monitored and updated and used as a basis for controlling costs. Annual budgets, operating and long-range plans will often disguise fundamental project-related problems, including maintenance costs which will only be highlighted by a life-cycle-cost review.

18.3 Plant selection and replacement

18.3.1 Plant-selection criteria

Perhaps the golden rule with plant selection is one of time—it is important not to be rushed into purchases of plant and equipment which subsequently prove to be inadequate to meet the needs of the manufacturing department. Careful selection will reap benefits for the company for many years to come.

The key to adequate plant selection is one of defining the requirements, in fact, drawing up the specification. Specification preparation is not a simple task to be completed in an odd afternoon. It is eventually to provide an absolute commitment between two parties to provide some equipment and/or service. It needs to be both detailed and yet clear and concise in its content in order that there can be no doubt about its meaning, and no room for misunderstanding. Inadequacies in design, performance, maintenance or operation can often be traced to inaccuracies or omissions in the original specification. It is essential that a dialogue should take place during the preparation of a specification between all interested functions, rather than leaving the task solely to one person to complete.

For plant selection the following general headings give a guide to the content of such a specification.

(1) *Purpose of the equipment*—this is the technical specification and the function which the plant is expected to serve.
(2) *The environment in which the plant is expected to operate*—is it to be in a normal workshop environment or is part of some external plant, possibly to be operated in tropical or sub-zero weather conditions?
(3) *Production capacity*—the throughput of the plant in terms of whatever its intended function is to be, e.g. the production output expected from it such as units per day, gallons per minute.
(4) *Availability expected*—this will provide some indication of the 'overcapacity' which will have to be provided in (3) to meet production demands. (See Section 18.1.6. for details of availability and related parameters.) At this point in time it is worth remembering that there are three shifts available in a day and it might possibly be a more economic proposition to operate the plant over two or more shifts, rather than confine it to a single day shift. This is often desirable if the capital cost involved is high compared with other parts of the plant in the production process.
(5) *Life of plant*—expressed in appropriate units, e.g. years, millions of litres.
(6) *Communications*—relevant authorities to be consulted, e.g. purchasing manager, manufacturing manager, maintenance manager, safety officer.
(7) *Capital spend available*—indicate the preferred and maximum levels. (Note that when the specification is forwarded to the supplier this section should be amended to read the agreed price.)
(8) *Acceptable on-going costs of operation and maintenance.*
(9) *Comparison with all aspects of similar equipment currently in use*—this is particularly important where plant is being upgraded.
(10) *Related equipment*—compatibility with other plant, especially important for process equipment.
(11) *Maintenance requirements*—supporting documentation, tools and test equipment, spare parts, etc.
(12) *Applicable standards*—full list of British or International Standards to which the new plant must comply.
(13) *Certification (acceptance conditions)*—details to be given of any reports or test certificates required prior to acceptance of the plant.
(14) *Life-cycle-cost summary*—the procedures outlined in Section 18.2 should be followed to produce a complete costing anticipated over the whole life of the plant.

18.3.2 Plant procurement

Having completed the specification of requirements, the next activity is to investigate the market place and select a suitable vendor. There are two distinct avenues of approach here, one of which will probably already have become apparent during the preparation of the plant specification. Is a standard range of equipment available to meet the need, possibly with some modifications? Alternatively, is a special unique piece of equipment required to be designed and built?

Whilst the user is always looking for improved performance and availability of his new equipment, he is unlikely to consider the full implications of maintenance and whether an investment in maintenance features would show benefits in cost of ownership. Unfortunately, this factor is rarely considered at the time of purchase of new equipment. Very often the maintenance manager is brought in late in the procurement cycle and sometimes only when the equipment arrives on the premises. The maintenance manager and his team have years of experience, which if correctly channelled could provide an investment in the future reduction of maintenance costs. Close cooperation with the equipment supplier, particularly in the early stages, provides a sound foundation for developing the best 'state of the art' technical approaches for maintenance.

In selecting the vendor, the company's previous experience with suppliers should be taken into account. It may be that the company has a vendor rating system in operation. If not an assessment may have to be made of an unknown supplier. In fact, it may well be necessary if a supplier who would normally supply a standard range of equipment is to be asked to make significant modifications or build a completely new piece of equipment to meet the specification.

Particular areas to bear in mind when selecting a vendor are:

(1) past performance or vendor rating;
(2) ability to meet the technical requirements of the specification;
(3) quality of equipment previously supplied and/or expected;
(4) delivery dates and the supplier's ability to meet them;
(5) price (value for money);
(6) on-going operational and maintenance costs of similar equipment produced by the vendor; and
(7) after-sales service available—this would include technical information available in the form of user and maintenance manuals, additional direct technical support available to resolve any problems which might subsequently arise, unique tools and spares.

Typical questions to be asked bfore finalising the purchase order for plant are as follows.

(1) Is the statement of requirements (the specification) adequate?
(2) Is the supplier fully conversant with these requirements and does he/she have the technical capability and production capacity to meet them?
(3) Will the plant be acceptable to operators in terms of safety and ergonomics?
(4) Will the plant meet the availability criteria defined—in terms of reliability and maintainability? Can the vendor provide test results and other data to support this requirement?
(5) Can the vendor provide the after sales service required, in particular spares support to meet the availability specified over the whole of the anticipated life of the plant? What are the inventory cost implications of this?
(6) Do the life-cycle costs meet the criteria originally established?

Before delivery of the plant, the purchaser should:

(1) establish effective lines of communication with the supplier, particularly if a long procurement cycle is involved;
(2) ensure that all the necessary services are available for installation and that the maintenance resources will be available;
(3) assess the requirement for trained personnel and ensure that adequate training of all staff concerned with the plant is undertaken, including operators and maintenance engineers; and
(4) ensure that all consumable items, on-site spares, tools and test equipment are available to meet expected demand.

18.3.3 After procurement of plant

There are a number of actions required by a purchaser of plant to complete the purchasing cycle. These are as follows.

(1) Check that all the acceptance conditions listed in the specification have been met.
(2) Ensure correct installation and commissioning. Many later problems can be traced back to hurried or inadequate activities at this stage.
(3) Ensure that the configuration of the plant is the same as that in the manuals and spare parts listing and that any modifications incorporated, or deviations allowed during manufacture, have been taken into account.
(4) Collect information on technical, maintenance or operational problems and maintenance and operational costs for future assessment of the next generation of plant. Feed this information back to the supplier as appropriate in order for him to improve the performance of subsequent plant for similar applications.

18.3.4 Plant replacement

In the life-cycle-cost examples described in Section 18.2, no account was taken of a replacement strategy, or the value which plant might have had in the second-hand market. Very often plant replacement is one of necessity for some reason or other, but using the life-cycle-cost process, it is possible to provide a justification for plant replacement on economic grounds. Strictly speaking, life-cycle costing can only be complete if a strategy for disposal and replacement is taken into consideration. It is often difficult to see a precise timing for disposal of equipment at the time of purchase, but there is usually some accounting practice or past history which will suggest a nominal life figure. Factors that will affect the disposal decision are:

(1) technological advances,
(2) change of requirement,
(3) reduced operating efficiency,
(4) increased maintenance costs,
(5) non-availability of spare parts, and
(6) prestige associated with new equipment.

Returning to the example of the vehicle (*Table 18.1*), and moving forward in time to the end of year 3, a new vehicle is now available at £40 000, with improved operating and maintenance costs. *Table 18.6* compares the options of purchasing a new vehicle and disposing of the older one for £10 000, against the alternative of continuing to run the older one to the end of year 6. As can be seen in this analysis, the total cost over the period under consideration is in favour of early replacement.

Table 18.6 Alternative replacement option

Year	*Capital* (×10³ £)	*Operating costs* (×10³ £)	*Maintenance costs* (×10³ £)	*Revenue losses* (×10³ £)	*Total (constant)* (×10³ £)	*Discount*	*Total (NPV)* (×10³ £)
3	40–(10)				30.0		30.0
4		3.5	2.0		5.5	0.91	5.0
5		3.5	2.0		5.5	0.83	4.6
6	(14)	3.5	2.0	1.5	(6.5)	0.75	(4.9)
Total							34.7
4		4	4.3	4	12.3	0.91	11.2
5		4	5.2	6	15.2	0.83	12.6
6	(2)	4	6.2	8	16.2	0.75	12.2
Total							36.0

18.4 Measurement of the effectiveness of maintenance

18.4.1 The need for measuring maintenance

Most companies know the direct costs of maintenance, but few companies really understand the opportunities associated with them. All too often the emphasis is either towards minimising the maintenance budget or repairing 'when it goes wrong', rather than taking the overall view of maintenance as a business opportunity.

The Asset Management Information Service (AMIS) is a Department of Trade and Industry supported scheme whereby a company can obtain assistance to conduct a self-audit of its maintenance activity. Some 450 of these audits have been carried out by MCP Management Consultants (the organisation funded to undertake these audits). They show that maintenance costs are typically between 1 and 8% of the annual sales turnover, the weighted average being 3.6%. These same surveys also point to relatively inefficient maintenance organisations. Through the scoring systems used in their audit they conclude that scores of around 50% are commonplace in areas described as 'general maintenance level', 'organisation and administration', 'work planning and control', 'cost control', 'productivity and maintenance effectiveness', 'materials management', 'failure analysis and betterment', 'training' and 'safety'. All these are key areas of the maintenance activity.

Repeat audits which were carried out 12–18 months after the original audit show that an improvement of 11–20% is quite achievable. This in turn demonstrates that significant improvements can be achieved in a relatively short time, providing key areas in which shortfalls are occurring can be identified, measured and monitored in such a way that corrective action can be taken to improve their performance.

The figures also indicate that 75% of the companies audited reported instances of unplanned work in excess of 20% of the time available, with over 35% having backlogs of work of between 1 and 5 weeks. Such figures do not augur well for the highest achievable measure of availability for their plant!

18.4.2 Maintenance costs

This is the traditional method of measuring maintenance, but whilst a cost figure without some qualification may be suitable as an overall budget or planning tool, it is extremely vulnerable to arbitrary manipulation by the accountants! The needs of the company accountant and the maintenance manager are very different. The former may not be concerned much beyond the total maintenance cost, whilst the latter needs a detailed breakdown of costs and their components to manage his day-to-day business. However, they do have one thing in common—that the totals must be compatible, and to achieve this it is important to ensure that no costs are omitted or double counted.

18.4.2.1 Maintenance hours

The maintenance engineers' time can be allocated to a number of different activities, and the individual maintenance manager will have to make the decision which to include in 'maintenance hours', but usually the term is applied to the actual work or 'hands-on' time spent by an individual engineer maintaining a piece of equipment.

18.4.2.2 Hourly rate

The engineers' time can be broken down into a minimum of the following categories:

(1) work time,
(2) travel time,
(3) administration time,
(4) waiting time, and
(5) non-available time (e.g. training, meetings, holidays, sickness).

All of the above contribute to the hourly rate, amounting to a utilisation factor as a direct overhead on the work time. Additional payments made for overtime or bonuses also contribute to this rate. Depending on the accounting convention used in an individual company, it may also be necessary to add an appropriate proportion of any of the following to the hourly rate as the indirect overhead:

(1) social and other benefits,
(2) training and documentation costs,
(3) capital costs, including tools and test equipment,
(4) vehicle running and other travel costs,
(5) clothing allowance,
(6) occupany costs,
(7) communication and on-going computer costs,
(8) other internal recharges, and
(9) outside services.

18.4.2.3 Parts cost

Parts can be internally manufactured items, repaired parts or bought in goods. Each type is likely to be recorded in the company accounts on a different basis.

If we start by considering the internally manufactured part, this has a value based on the materials and manpower required to produce it. It will also have had a proportion of overhead added to it, in exactly the same way as the overheads were added to the maintenance engineers' time. It may be held in inventory at this price, or perhaps a mark-up will also have been included in the transfer price between the manufacturing division and the stores department. This will depend on internal company accounting policies and organisation, principally whether the manufacturing division operates as a profit centre. Again the opportunity exists for a mark-up to be added when the stores issue a part. This time it would be to cover the overheads associated with holding the part in inventory together with any profit mark-up, if appropriate. Consequently, a part could easily be valued at double its production cost by the time it reaches the maintenance engineer.

The repaired part is more difficult to cost, depending on where it was repaired. At the time of removal from an equipment as a failed part it has zero value, being completely useless, and if there was no policy to repair it, it would be thrown away. If the maintenance engineer can repair it to its full working specification (and this implies that its life expectancy would be the same as that of a new part), then the part is available for use by the engineer at the sum of time and materials costs which were used in the repair. Care must be taken in accounting for this in recording maintenance costs.

If the repaired part is sent away from the maintenance function for repair, then it is usual for the inventory department to allow a credit to maintenance, and when this part is reissued to them as fully serviceable, it is accepted at the new part price.

The bought-in part has a price tag on arrival and it often depends on how far back down the chain of maintenance, stores and manufacturing that the purchasing function is located, as to the final cost to maintenance.

18.4.3 Maintenance indices

A number of different indices are used to measure maintainability and maintenance and, before discussing these in detail, a few general comments are appropriate.

The first is the data source. Any figures which are used in the compilation of an index must be compatible and measured on the same basis if any meaningful comparison is to be made. It follows, therefore, that the most effective use of an index is for the purpose of tracking from one point to another, using the same database, to determine the direction of trends. However, having said that, further notes are given below on each index. It is also essential that an appropriate index is chosen to suit a particular purpose. This is demonstrated in *Table 18.7* and the accompanying narrative.

Availability has already been defined in Section 18.1.6 as:

$$\text{Availability} = \frac{\text{Possible operating time} - \text{Down time}}{\text{Possible operating time}}$$

Maintenance effectiveness can be used as a measure of maintainability, and is a useful basis for comparing different equipment both from design projections and from historical data:

$$\text{Maintenance effectiveness} = \frac{\text{Actual equipment use}}{\text{Maintenance effort required}}$$

Units should be chosen to suit the working environment of the equipment and a typical parameter would be 'operating hours per maintenance hour'. The maintenance effort would be dependent on reliability and mean time to repair. The larger the resulting maintenance-effectiveness number, the better the maintainability.

In *Table 18.7* a comparison is made of the performance of four different equipment options which would perform a similar function, and which a user might consider purchasing. The various indices are calculated and a comparison can be made between them from a maintenance point of view. It will be seen that, despite the quite different failure rates expected, there is only one unit difference in availability over the range of the four options listed. It will also be seen that the maintenance effectiveness and maintenance cost figures are almost in the reverse order. Perhaps the figures used in the example are a little more extreme than would be found in practice, but it does demonstrate the variations found by using different indices.

The *prime maintenance index* is another factor often used for comparing one piece of equipment with another prior to purchase, and is favoured by accountants, but may be misleading from the maintenance manager's viewpoint.

$$\text{Prime maintenance index} = \frac{\text{Current annual maintenance costs}}{\text{Capital replacement cost}}$$

This index should be compared with the formula for maintenance effectiveness which performs the same function and is much more relevant to the maintenance activity. The significant difference between the two is that the maintenance effectiveness is related to usage of the equipment, whereas the prime maintenance index is simply a cost comparison. In the latter case the index can be influenced by outside factors such as purchasing discounts, equipment price changes and taxation considerations.

18.5 Operational aspects of maintenance

18.5.1 Maintenance planning and control

Many aspects of maintenance planning and control will be predetermined by company organisation and practices. There is no 'right' way which is applicable to every situation. However, of prime importance is that maintenance as a function is represented at a sufficiently high level within the

Table 18.7 Example showing a comparison of indices

Option	*Failures per 1000 h*	*Average repair cost* (£)	*Average repair time* (h)	*Availability* (%)	*Maintenance effectiveness*	*Maintenance cost per 1000 h* (£)
A	7	150	0.7	99.5	204	1050
B	10	100	1.0	99.0	100	1000
C	2	250	5.0	99.0	100	500
D	15	50	1.0	98.5	67	750

company to contribute to the overall profitability of the company.

The key to efficient use of maintenance resources is the work-control function. Ideally, this should be a centralised function across all maintenance resources, unless the maintenance function itself is extremely large, very diverse in trade skills or geographically widespread. In other words, its strength lies in the ability of the work controller to be able quickly to switch manpower from one activity to another according to the priorities of the moment. The work controller's function is two-fold:

(1) matching of the workload with resources, and
(2) allocate priorities within this activity.

It is important that the work controller has the authority to undertake these activities without recourse to higher management, other than in exceptional situations.

The Work Controller's tasks are as follows:

(1) receive requests for maintenance;
(2) check priority of request (immediate, within 2 hours, etc.);
(3) record relevant information regarding the request, including time received;
(4) allocate the job to the maintenance engineer or team (having the appropriate skills) and prioritise their activities;
(5) record details of work done, spares used, etc., as advised by the maintenance engineer and note the time that the work was completed; and
(6) update records as appropriate.

In its simplest form this may be done manually on record cards, or, as is more usual these days, details will be entered on to a computer terminal as they are received.

Inevitably, from time to time there will be insufficient manpower to meet all the immediate demands and the earlier a pending *overload situation* can be identified, the more chance there will be of taking remedial action sufficiently early to avoid any loss of availability of the equipment under the care of the maintenance department. Whether manual or computerised, the general approach to identifying and resolving such a situation will be the same.

When a request for maintenance is received, the work controller identifies the equipment and the maintenance engineer of the appropriate skill (e.g. electrician, mechanical fitter). The time is noted together with the maximum acceptable time for the engineer to arrive on site and the anticipated work time to resolve the problem. At any time, therefore, a manual or computerised record is available for each pending job in terms of the number of hours of work anticipated and the time frame in which the job should be carried out. If this is then matched to the availability of the maintenance engineers, any overload situation can be quickly identified. The task is then to eliminate the overload either by rescheduling non-urgent work or by bringing in extra manpower from another part of the plant which, hopefully, will not be overloaded at the same time.

18.5.2 Computers and maintenance

There are a number of suppliers of commercial software packages who can provide very comprehensive maintenance management systems which will incorporate this work queuing facility as well as many other features from plant records to stock control.

Any computer system must satisfy two fundamental requirements:

(1) its cost of implementation and operation must be offset by savings in direct resources or increased efficiency; and
(2) the data output must be as complete, timely and accurate as the manual system it is replacing.

An investigation into existing systems will identify a number of key factors such as:

(1) data sources, their completeness and accuracy;
(2) reports produced;
(3) capital and personnel resources currently in use;
(4) effectiveness of the system; and
(5) personnel involved in receiving, assessing and making decisions on the data.

This should reveal the main characteristics of the current system, its shortfalls and redundancies, enabling a detailed specification for a new system to be produced.

At this stage, in selecting a computer system, 100% reporting of everything directly by the maintenance engineer is possible—at a cost! Depending on the quantity and variation of equipment being maintained, together with the frequency of maintenance activities, can 5 min terminal input by the maintenance engineer be justified after every call-out? If it can (and the extra capital cost can also be justified), then it may be worth considering installing terminals for use of the maintenance engineer. There may be other advantages in doing this, such as the calling up of job routines on the computer screen and the ordering of parts required direct to and from the shop floor.

It is always preferable to identify the software first before considering the purchase of any hardware. One particular area which should be given close attention is the ability of the software to cope adequately with numbers from the existing part numbering system used by the maintenance function. It is not just the numbers themselves which may cause a problem, but the ability of the software system to sort them in a meaningful manner, e.g. differentiating between piece parts, purchased parts (possibly from different suppliers with their own unique numbering systems) and major assemblies.

If technical reporting coverage is to be included in the system, it will also be necessary for the system to be able to produce special reports such as reliability growth curves, both for specific parts (or groups of parts) and problems. The system also needs to be capable of batching reporting codes together to produce an historical database in order to progress identified problems and their solutions.

18.5.3 The profit-centre concept

Traditionally, maintenance has been regarded as an expense. If equipment is broken down then it will inevitably cost money to return it to its fully operational condition again. That is the generally accepted view of maintenance. Are there situations where it is not just an expense? Are there ways of offsetting the costs of maintenance? Can it be shown to make a profit? The answer to all three questions is 'yes', but it can only be achieved by a company-wide commitment to make profit an objective of the maintenance department.

Part of this commitment must include the ability of the accounting system to accept recharges between departments, and the introduction of transfer pricing policies for goods and services within the organisation. The latter will be necessary to determine the separate performances of both the 'selling' and 'buying' departments. This procedure may not be practicable for the smaller company, and care should be taken to ensure that any additional accounting practices are not more expensive to operate than the estimated overall profits that are likely to ensue from introducing the concept. It must be

remembered that profits on internal transfers do not in themselves contribute to those of the company. In addition, if a manager decides to make a purchase outside the company in preference to purchasing from an internal department, he may unintentionally create underutilisation in the 'selling' department, thereby increasing the departmental overhead and reducing the overall profitability of the company.

There are, however, a number of intangible benefits to such a system including improved motivation, increased efficiency, better decision-making and the decentralisation of many management decisions.

The simple formula on which all businesses exist is:

Income = Costs + Profit

It follows, therefore, that to increase the profit, either the costs must be decreased or the income increased. There are clearly two opportunities for the maintenance function.

18.5.4 Direct and indirect costs of maintenance

Consider first the costs of maintenance. Part of the maintenance manager's job specification will inevitably be to minimise maintenance costs. So why not recognise this as an internal profit? If the role of the maintenance manager is compared with that of the production or sales manager, his contribution is equally as valuable to the overall profit of the company.

Improvements in sales means more production and some of that could have been brought about by improved availability of manufacturing equipment. Some of the reduction in production costs could also have been attributed to the improved availability. So, providing these gains can be isolated, they can be credited to the correct function, namely maintenance.

In Section 18.4.2 the constituent elements of maintenance cost were discussed and direct costs and indirect overheads were identified. The latter should not be confused with indirect costs, which are charges incurred (or in income terms, perhaps income derived) by other departments as a direct consequence of maintenance activity.

18.5.4.1 Direct costs of maintenance

These can usually be determined at the time of purchase of the equipment using the life-cycle-costing process, and the maintenance department would not be expected to deviate from these estimated costs without some reason being identified. It may be that the equipment does not perform as intended, perhaps it is not so easy to maintain as the manufacturer originally advised, or maybe its utilisation has increased over that which was planned at the time of purchase. Any of these reasons could justify additional direct cost being incurred by the maintenance department. On the other hand, less resources might be needed for the opposite reasons. These are changes to plan and should be treated separately from departmental profit and loss.

However, if genuine efficiencies are introduced, either in manpower or parts (either usage or stock holding, if the latter is part of the maintenance department budget), then these will lead to a reduction in direct maintenance costs. This in turn will contribute directly to the overall profit of the company, providing that the departmental targets for maintenance are being met and that undetected costs are not being incurred in other departments.

18.5.4.2 Indirect costs of maintenance

These arise in other departments of the business. In Section 18.2, for example, when considering life-cycle costs for the purchase of a new vehicle, there was the opportunity to trade-off operating costs, maintenance costs and revenue losses to achieve the lowest overall cost. This was essentially a planning exercise done before any decision was made to purchase. Now, supposing that the overhaul option detailed in *Table 18.2* was not considered as part of the original purchase planning process, but was raised as an option by the maintenance department in year 4, then this overall profit to the company of £3600 would be achieved largely by eliminating expected revenue losses due to increased availability of the vehicle.

Other areas in which similar profits may be identified include:

(1) lowering of reject rates by improving maintenance;
(2) elimination of overcapacity planning by departments allowing for extra maintenance downtime; and
(3) reduction in support to maintenance in other departments due to more efficient maintenance.

Figure 18.11 shows a summary of the principal categories attributable to the cost of maintenance, which would provide opportunities for profit identification. Note that this is a positive approach to profit attributable to maintenance, rather than a cost-cutting exercise.

18.5.5 Maintenance revenue

Looking now at the 'income' side of the equation, maintenance can only contribute additional revenue if it can widen the range of services which it can provide to the user, and be able to charge for these services.

Depending on the environment in which the maintenance department operates, maintenance revenue may be derived from both within the company and from other quite separate organisations (see *Figure 18.12*). However, in a manufacturing environment it will probably be only appropriate to consider those opportunities which may exist within the company.

The first source of revenue is, therefore, to charge for maintenance. Internally, within a company, this may be unnecessarily cumbersome, particularly on a time and materials basis. However, there are merits in building the departmental budget in the form of a number of maintenance contracts to be agreed with the user departments (see Section 18.5.7). One of the principal advantages of this approach is that the maintenance manager has a clear plan of the number and variation of equipment that he/she will be expected to

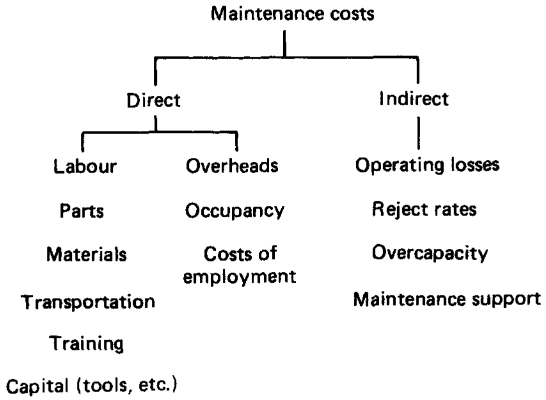


Figure 18.11 The costs of maintenance

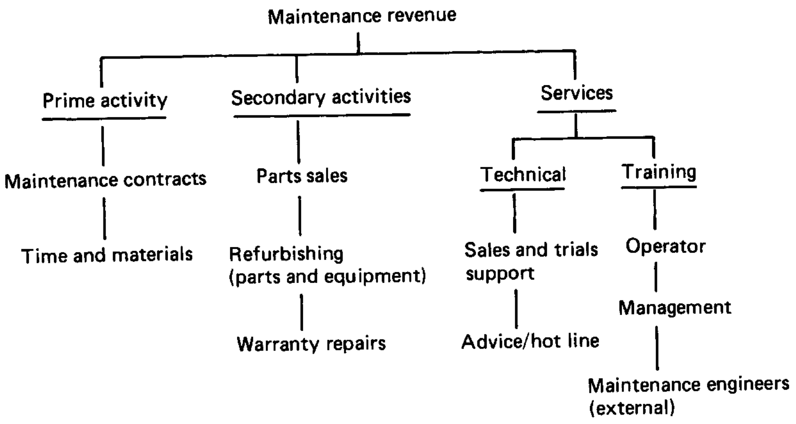


Figure 18.12 Sources of maintenance revenue

maintain during the period under review. Any additional requirements can then be added as a result of negotiation rather than surprise!

The second source is to increase the services offered by the maintenance department. It should be borne in mind that the prime objective of the maintenance department is to keep the equipment with which it is charged in fully operational order. It would be unwise to extend the services of the department to the extent that its activities detracted too much from this prime objective, but a review could be undertaken to identify complementary activities which would provide a more complete service to the user.

Typical complementary activities would include the following.

(1) *Training*—user or operator training in the correct use of the equipment, possibly extending to some elementary maintenance activities. There may also be a requirement to familiarise management and technical staff with the equipment.
(2) *Support for internal and external exhibitions*—any special equipment trials which might be undertaken, together with technical advice could also be considered.
(3) *Resiting and recommissioning*—this activity is often absorbed in the day-to-day operation of the maintenance department, but really should be recharged to the user department who require the equipment to be moved to its new location. When required, site surveys could all be undertaken as part of this activity.
(4) *Refurbishing of spares and equipment*—some refurbishing of spares may be undertaken by the maintenance department as part of its own economic approach to maintenance, in which case there would be no other department that could be recharged. However, where this activity can be performed for other equipment not normally maintained within the company, then the opportunity for recharging exists.

18.5.6 Pricing of maintenance

Recharges for services have been discussed above, but without any indication as to how these should be priced. A strategy on which recharges can be based should meet the following objectives.

(1) The price for maintenance activity should be consistent with the overall objectives of the company. For internal maintenance they should be at such a level that does not encourage the use of external sources.
(2) The price level should be such that an accurate assessment can be made of the performance of the maintenance department, in particular with respect to its profit contribution to the company.
(3) The price level should be such as to ensure that the benefits of the profit centre concept are achieved, if possible independent of the action of other functions within the company.

There are three ways in which the price for maintenance can be set:

(1) by basing the level at a competitive market price,
(2) by cost-based pricing, and
(3) by an agreed price.

The preferred method must always be the first in the list above, but at times this may be difficult to establish. Cost-based pricing, using accepted labour rates and parts costs is a good alternative, and is often the preferred method for establishing internal recharges. The 'agreed' price usually forms the basis for the maintenance contract.

Maintenance costs are closely defined within an organisation, but maintenance pricing also depends on the level of service provided to the user. Manufacturing departments may be prepared to accept charges at a higher rate for shorter response times to their requests for maintenance on their key items of equipment, or the provision of maintenance cover outside the normal working day.

The allocation of overheads on maintenance costs should be examined. It is traditional to relate company overheads to man-hours within a departmental budget, but this may not be the most effective method from the point of view of pricing and establishing a satisfactory profit. It may well be possible to absorb a higher level of overhead on the maintenance of some equipment than others, to obtain a more competitive rate. Some alternative approaches to this are shown in *Table 18.8*, where the maintenance of three different equipment ranges is considered. As can be seen, if the pricing of maintenance for equipment C is not competitive when the overhead is based on labour-hours, this improves dramatically when the allocation is based on the number of pieces of equipment maintained, but severely penalises A. Two other examples show the effect of basing the allocation on call-outs and a random weighting by product range, the latter benefiting both A and C with a relatively small percentage impact on B.

Table 18.8 Example showing options for overhead allocation

	Equipment range		
	A	*B*	*C*
No. of equipments maintained	500	400	200
Call-outs/equipment/year	1	4	15
Average hours/call-out	1.0	1.5	2.0
Labour revenue/equipment (excluding overheads)	£20	£120	£600
Total overhead allocation based on labour-hours	£5000	£24 000	£60 000
Annual revenue options/equipment:			
Based on labour-hours	£30	£180	£900
Based on No. of equipments maintained	£101	£201	£681
Based on No. of call-outs	£38	£190	£862
Based on weighting by equipment range, 1:20:25	£24	£217	£842

18.5.7 Maintenance contracts

There are advantages to both the user and the maintenance function in obtaining agreement for maintenance on a fixed price contract basis. Both organisations have an agreed commitment (to pay and provide maintenance), and know exactly the conditions of this agreement and can plan for it with confidence. The user is assured of regular preventive maintenance (if such a scheme exists for the particular equipment), speedy response to any call for assistance, and a minimum of paperwork. From the point of view of the maintenance manager he has a known range of equipment to service at a fixed revenue, with only the need to raise regular user invoices, rather than one for each maintenance request.

However the basic price of a maintenance contract is established, it should be structured in the contract in such a way that it can be varied according to the equipment utilisation. This can either be by pricing it in relation to a meter reading (for example, £1.00 for every hour of operation), or by including an option to amend the charge should the utilisation vary beyond certain agreed limits. Above all, maintenance contracts must be competitive in the market place and provide an adequate rate of return. If a satisfactory rate of return at a reasonable market price is not achievable, then the maintenance manager will need to review his cost base. Some areas for examination are:

(1) Is the level of service (response time, labour skills, parts back-up and any other commitments to the user) higher than it need be?
(2) Is the overhead structure unfairly loading the cost base? Can it be divided any other way or spread more widely to produce a lower rate?
(3) Could the balance of the level of service offered to other departments or equipment users be varied to justify a differential charging rate (that is a higher rate for some special service)?

Failing that, there is little option other than to either accept a lower rate of return (perhaps only from some users) or decline to offer the maintenance contract. The latter course of action suggests an underfunded maintenance department!

18.5.8 Modifications

One method of reducing maintenance costs, and hence of improving profitability, which has not yet been discussed, is to improve the performance of the equipment. As has been said earlier, reliability and maintainability can only be designed into a piece of equipment, but designs can be changed and modifications introduced to improve either of these features.

The timing of such modifications is, however, very important. There is often a considerable lead time involved between identifying the need, designing the change, producing the hardware and instructions for fitting, and scheduling the work of modifying the equipment. Consequently, however attractive a modification looks at the time of its inception, an economic study needs to be undertaken, taking into account the lead time involved to fitting and the expected life of the equipment, to ensure that there is adequate payback to make it all worthwhile. If the parts for the modification must be purchased by the user, and particularly if freight and duty costs are involved, then this is another negative aspect to be taken into account.

Consider again the vehicle which was used as an example in Section 18.2.2 (*Table 18.1*). Supposing a gain of £300 per annum in operating costs is to be had for an outlay of parts at £600 and labour for fitting of £100. If this decision was taken in year 3, then at first glance an outlay of a total of £700 to gain £300 per year looks reasonably attractive. Supposing it will be early in year 4 before the modification is actually fitted, then this still leaves nearly 3 years to achieve the benefits. However, the marginal analysis shown in *Table 18.9* demonstrates that, if everything goes according to plan, the best that can be achieved is to break even. It is therefore probably not a worthwhile investment.

18.5.9 Refurbishing of parts and equipment

Refurbishing can be a useful source of profit to the maintenance department and, at the same time, can improve the overall utilisation of the staff. This can be achieved by utilising the 'waiting' time normally experienced by maintenance engineers. It is a useful buffer activity in terms of labour hours and, if successful, can often justify additional staff, either dedicated to refurbishing or simply to increase the overall resource level available for both maintenance and refurbishing activities. At the same time, having the additional staff may improve the response time to the calls for assistance from the user. However, it is important that the quality of the refurbished parts is of a high standard, as for all practical purposes a refurbished part should have an equivalent performance to that of a new part. If it does not have the same quality, then

Table 18.9 Cost and benefits of a modification

Year	Original case				Modified case			
	Operating cost (×10^3 £)	*Maintenance cost* (×10^3 £)	*Discount*	*Total (NPV)* (×10^3 £)	*Operating cost* (×10^3 £)	*Maintenance cost* (×10^3 £)	*Parts cost* (×10^3 £)	*Total (NPV)* (×10^3 £)
3	2.8	3.6	0.91	6.4	2.8	3.6		6.4
4	2.8	4.3	0.83	6.5	2.6	4.4	0.6	6.9
5	2.8	5.2	0.75	6.6	2.5	5.2		6.4
6	2.8	6.2		6.7	2.5	6.2		6.5
Total				26.2				26.2

this must be taken into account when assessing the success or otherwise of the refurbishing activity. For this reason, all refurbished parts should carry some identification to indicate their past history.

The income from refurbishing a part is equivalent to the price of a new part, less the cost of refurbishing, less any credit which might normally be available from the original supplier or for scrap. The credit possibility should not be overlooked because it might make a refurbishing opportunity non-viable.

In the example shown below, nominal credit has been allowed and added to an assumed revenue derived from other sources, the profit and loss account prior to establishing a refurbishing operation being as follows:

		×10^3£
Maintenance revenue		8500
Credit for parts x		100
Total income		8600
Maintenance costs	Labour	4800
	Parts x	400
	Other parts	900
Fixed overhead		2400
Total costs		8500
Profit		8600 − 8500
		= 100

After establishing the refurbishing operation, the account will appear as follows. Note the division of the fixed overhead at 50% of labour costs, assuming no additional staff are needed to undertake the new activity.

		×10^3 £
Maintenance revenue		8600
Refurbishing costs	Labour	300
	Parts	50
	Fixed overhead	150
Total refurbishing costs (parts x)		500
Maintenance costs	Labour	4500
	Parts x	500
	Other parts	900
Fixed overhead		2250
Total costs		8150
Profit		8600 − 8150
		= 450

It can be seen that the refurbishing activity by itself is a loss-making activity, since parts x are refurbished at £500 000, whereas they could be purchased for £300 000 allowing for the credit. In spite of that the overall profit situation of the maintenance activity increases from £100 000 to £450 000! This increase would be sufficient to absorb some additional staff to assist with the maintenance activity, the numbers and resulting profit depending on how the labour-related fixed overhead increases.

This type of costing exercise demonstrates that significant savings are achievable by better utilisation of staff, which might not have been apparent unless a profit-centre concept for maintenance had been introduced.

18.6 Preventive maintenance

18.6.1 Classifying types of maintenance

There are two basic types of maintenance: the scheduled and the unscheduled. For maximum availability of the equipment or plant the unscheduled maintenance should be kept to a minimum. However, to some extent this minimising of maintenance needs to be tempered with the economic consequences of 'overmaintaining'. In a real situation, therefore, unscheduled maintenance will inevitably occur.

Where failures can be predicted, then maintenance can be scheduled. The origins of data for scheduled maintenance can be derived from a number of different sources, such as:

(1) safety requirements,
(2) manufacturer's recommended maintenance routines,
(3) operating experience,
(4) condition monitoring,
(5) non-essential performance deterioration, and
(6) non-essential equipment failures.

The first of these categories, safety requirements, is typified by a crane which would require weekly inspections and periodic testing of the wire rope. The next category is similar, but is a manufacturers's recommendation, rather than a mandatory requirement. It is usually related to recommended time intervals and includes such activities as lubrication, inspection, testing and replacing of some parts which have a short life. With experience these schedules are often added to or amended by the maintenance department themselves, producing a listing of activities to be performed at regular intervals. The last two categories in the list above are those items which will not affect the availability of the equipment in question, or perhaps allow it to continue to operate with some degraded aspect of performance which can be tolerated for a short time, until the next opportunity to carry out preventive maintenance.

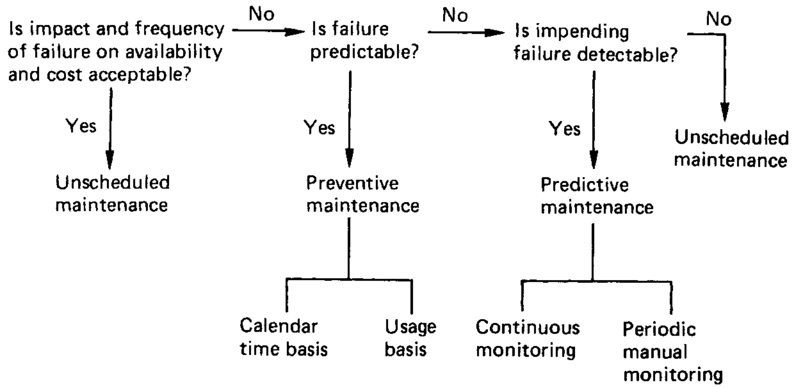


Figure 18.13 Decision-tree approach to determining the optimum maintenance period

Condition monitoring can be very effective where a failure can be anticipated by some form of deteriorating performance over a period of time to allow the maintenance engineer to take remedial action. This process is known as predictive maintenance and is dealt with in Section 18.7.

Therefore, the maintenance categories may be summarised as:

Maintenance = Unscheduled + Preventive + Predictive

The guiding principles are:

(1) 'as little as possible, as often as necessary', and
(2) 'maximise availability economically'.

Figure 18.13 shows a decision-tree approach to determining the optimum form of maintenance. This approach can be used for any particular application.

18.6.2 Failure patterns

Failures happen in many different ways and for many different reasons, not the least of which are during transport and storage. These are completely random and each should be treated as an individual problem, although obviously similar problems point to a need for investigation and correction. For operating failures it has been generally accepted for many years that the familiar bath-tub curve (*Figure 18.14*) represents a typical failure pattern.

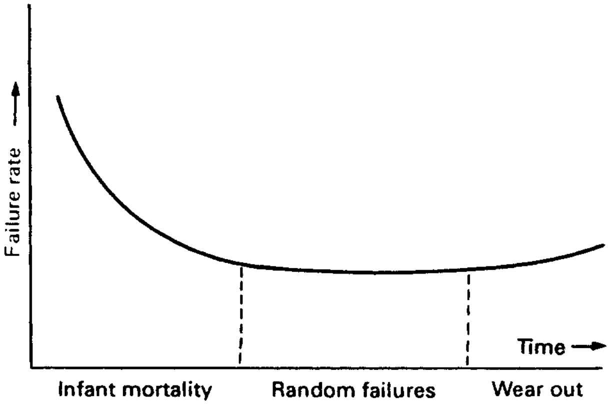


Figure 18.14 Typical failure pattern

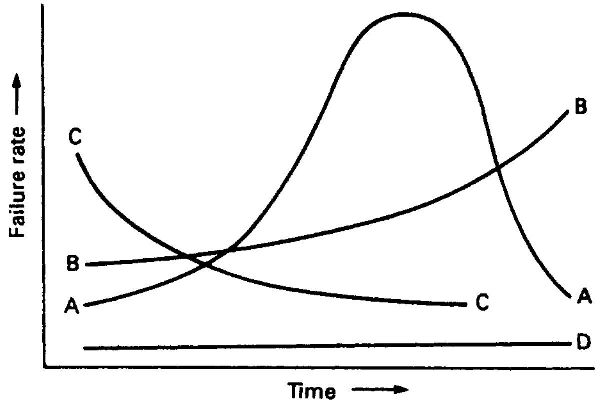


Figure 18.15 Different forms of failure pattern

In a well designed and manufactured equipment, the majority of failures are completely random. *Random failures* would occur in the bath-tub curve shown above during the period when the failure rate is at its lowest and, for most purposes, can be regarded as constant.

Other common forms of failure pattern are shown in *Figure 18.15*. Line A represents a wear-out failure for an item that has a finite life, such as a light bulb which is guaranteed for 1000 h, but the majority of which will fail before double that life is reached. Line B shows an item which has a failure pattern which is essentially the second two sections of the bath-tub curve, differing in as much as it does not exhibit any significant infant mortality. This type of curve is generally true for many mechanical moving parts which are subject to gradual wear, such as a bearing. Line C demonstrates the infant mortality pattern and is effectively the first section of the bath-tub curve. The failure of electronic components generally is typified by this curve. Line D can usually be applied to represent parts which seldom fail, or are subject to damage or loss during use or maintenance.

18.6.3 Establishing preventive-maintenance routines

The ideal contenders for inclusion in a preventive maintenance programme are those items which have a performance characteristic which falls below an acceptable level in a

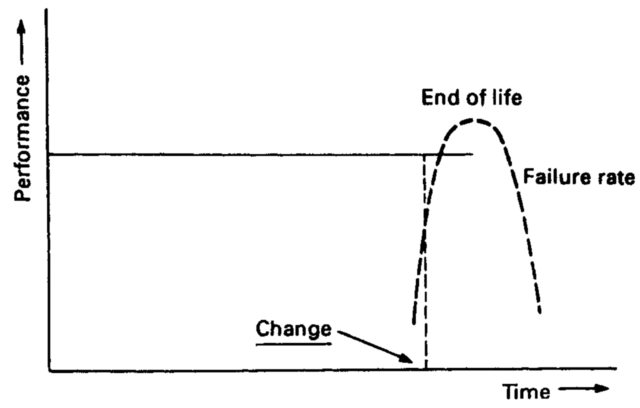


Figure 18.16 Preventive maintenance for a part showing instantaneous failure

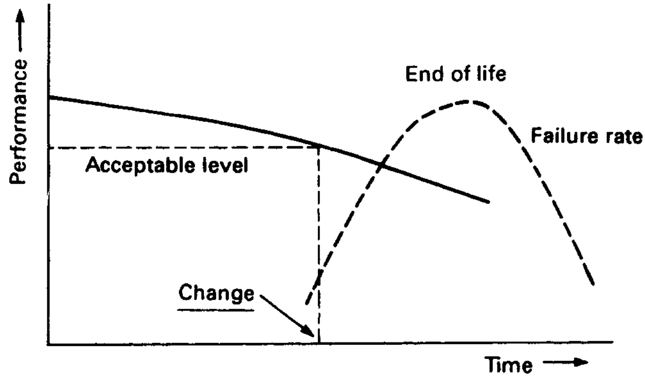


Figure 18.17 Preventive maintenance for a part showing degraded performance

consistent interval of time. In this instance the word 'time' may be used to indicate elapsed time, operating time or an output quantity or usage. In fact, any way that can be measured relative to the equipment. Two types of performance degradation are common, and these are illustrated in *Figures 18.16* and *18.17*.

Figure 18.16 shows a type of failure that was discussed earlier in the previous section, such as the lamp bulb which either lights or does not, with failure which is instantaneous after perhaps 1000 h. *Figure 18.17* demonstrates another type of failure where the performance gradually degenerates over a period of time until it becomes unacceptable. Both graphs also show the distribution of failures with time and an economic change point can be determined statistically from the failure pattern (or design specification if no historical data exist).

The method of actually determining the most economic preventive maintenance interval is best demonstrated by an example. Suppose when an unscheduled maintenance activity is made, a part costing £5 is replaced at a total labour cost of £20 and production to the value of £100 is lost. The failure pattern of the part follows a normal distribution curve, similar to that shown in *Figure 18.16*. The impact of alternative

Table 18.10 Impact of alternative preventive maintenance policies

Proposed PM interval (h)	*Cumulative % parts failed*	*Costs per 1000 h* (£)		
		PM	*UM*	*Total*
900	0.5	28	1	29
1000	3	24	4	28
1100	14	20	16	36
1200	36	14	38	52
1700	99	—	73	73

PM, preventive maintenance; UM, unscheduled maintenance.

preventive maintenance policies is shown in *Table 18.10*. It is assumed that preventive maintenance can be undertaken at a convenient time when there will not be any production losses. As can be seen, an interval of around 1000 h would be the most economic approach.

The costs given in the table are simplified in as much as all failures and replacements were calculated from time zero, without taking into account the effect of the clock resetting every time a part is replaced. This more complicated situation may be analysed by using computer based simulation programmes. This involves simulating failures, statistically distributed with the expected probability distribution function, following the logic of the proposed preventive maintenance strategy, and calculating the resulting costs. By carrying this process through a large number of failures and replacements, typically 100 or more, average costs can be obtained using the different strategies, and the optimum choice selected. This method of selection is known as a 'Monte Carlo simulation'.

That a part fails at a predicted time is no certainty that it is a suitable candidate to be incorporated in a preventative-maintenance programme. The effect of the failure is the significant factor. Does it cause a complete or partial shutdown of the equipment? Is the failure rate pattern sufficiently consistent to be able to predict an economic change point? When it fails, does it cause a safety hazard? The answer to questions such as these will determine the need and the appropriate interval for preventive maintenance.

Assuming that the replacement of individual critical parts determines an interval for preventive maintenance, then it will be economic to schedule other activities at the same time, such as lubrication. It may also be an economic proposition to bring forward other activities to align them with the primary preventive maintenance timing, so that the bulk of preventive maintenance work can be done at the same time.

18.7 Condition monitoring

18.7.1 Introduction to condition monitoring

Condition monitoring is a method of providing maintenance management with a means of determining the technical state of their machines or equipment on an on-going basis. This basic concept is known as 'predictive maintenance' (see Section 18.6.1), and its purpose is to anticipate when a breakdown is likely to occur. This approach involves some form of monitoring which could vary from the simple visual observation of dirt or smoke, the feeling of heat in a critical bearing housing, for example, to the highly sophisticated computer monitoring system.

There are three techniques generally used in condition based maintenance.

(1) *Inspection*—periodic checks on any measurable parameter where a degrading condition may be detected by the human senses.
(2) *Comparison*—routine checks of a parameter which must fall between certain specified limits.
(3) *Trend*—measurement of parameters recorded at set intervals and plotted to detect a changing situation.

The first technique involves traditional methods adopted by experienced maintenance engineers and relies entirely on the human senses such as feeling, hearing and seeing. None of these checks is entirely reliable or repeatable in modern industry. Perhaps the nearest check to this is the coin standing on edge on top of a bearing housing, or a similar piece of equipment.

In the previous section on preventive maintenance, *Figures 18.16* and *18.17* show two examples of different failure characteristics. In the first figure there is a sudden end to the working life of a component, and in the other there is a gradual degradation of performance. If this degradation is fully predictable over the life of the component, then replacement as part of a preventive-maintenance schedule will be the most economic. If, however, the failure might occur at an unpredictable time, through wear or a change in the usage pattern, for example, then it is an ideal candidate for a modern condition-monitoring installation.

It is important to realise that condition monitoring can be approached using one or both of two alternative strategies. Firstly, as a continuous process on line with the equipment; or secondly, by monitoring at set intervals either by using hand held portable equipment or by periodic scanning of fixed sensors from a central monitoring facility.

A good example of condition monitoring and its wide opportunities is a coolant system, where there are four parameters which may be very easily monitored. The first is the pressure at the pump outlet, the second is the vibration levels of the pump bearings, the third is the coolant level in the reservoir, and the fourth is the coolant temperature. Obviously, in practice it would not be necessary to monitor all these parameters, and in all probability two would suffice.

Today, manufacturing is becoming more and more capital intensive and with the introduction of such techniques as just-in-time (JIT), it is vitally important that equipment is kept operational. Hence reliable, detailed and readily available information on machine condition enables the maintenance manager to ensure that downtime during operational hours is kept to the lowest possible level.

18.7.2 Alternative condition-monitoring strategies

The first and most important task when considering introducing condition monitoring is to undertake a thorough survey of the maintenance activities and the likely breakdown characteristics of the equipment. With this knowledge the benefits of the introduction of a condition-monitoring programme may be assessed. There is an up-front cost in introducing these techniques, in planning time and monitoring equipment (sensors and data analysers) which should not be underestimated.

The first decision to be made in the selection of a condition-monitoring system is the choice of parameters to be measured and the frequency of monitoring. It is probably self-evident that the more frequently the equipment is monitored, the more expensive is the monitoring installation. However, a careful cost analysis may well show considerable longer term benefits in a high initial cost of fully automated monitoring equipment.

The alternatives available are reviewed below.

18.7.2.1 Automatic monitoring

This is the process of continuous on-line surveillance and interpretation of machine performance. It inevitably involves the setting up of a computerised monitoring facility with appropriate software to interpret the data being received from machine-mounted monitoring sensors. The characteristics of the equipment being monitored will be instantly transferred to a database, providing the maintenance engineer with immediate trend data from which a pending malfunction may be quickly and automatically identified.

18.7.2.2 Periodic monitoring

The alternative approach to continuous surveillance is periodic monitoring. As has been mentioned above, this may take two forms.

Automatic periodic monitoring This is essentially a simpler form of continuous surveillance, with the difference that the fixed monitoring points on the machine are interrogated at predetermined intervals from a single scanning installation. The intervals will depend on a combination of the likelihood of failure and the effect of the failure, that is to say how critical a failure would be to the overall performance of the plant. The advantage of such a system is that it does not involve the same amount of centralised computer hardware as the continuous-surveillance method.

Manual periodic monitoring This is a less capital-intensive approach, but does utilise a higher amount of labour. The procedure requires the use of a portable data collector, hand held by the maintenance technician, which is taken around from unit to unit to analyse the characteristics of individual parts of the machinery. Particular care is necessary in selecting monitoring points to ensure good repeatability and accuracy of the readings taken.

(1) The location of the monitoring point must be such that the pick-up module or probe can be quickly and simply located in its measuring position.
(2) The output from the monitored signals must be such that they are independent of the manner in which the pick-up module is attached in its measuring position.

One of the most commonly used methods of attachment of pick-up modules is to use a magnetic base (although this is unlikely to be suitable for any measurement involving eddy currents).

The results from such monitoring are continually plotted, either manually or through a computerised system, and any deviation from one reading to the next can be quickly detected and any appropriate corrective action taken.

The frequency of measurement must be determined in the light of experience with similar types of equipment, but a general guide is given in *Table 18.11*.

In order to maximise the efficiency of the measuring procedure, machines and measuring points should be grouped into 'routes' with regard to their measuring cycles, measurement equipment and geographical location. If a portable data collector is used, then this may be preprogrammed to provide the most cost-effective measuring sequence.

18.7.3 Condition-monitoring applications and sensing equipment

Depending on the design of the machine to be monitored, one or more parameters can be measured. Different types of

Table 18.11 Typical periodic monitoring routines

Frequency	*Equipment category*
	Critical to production. Operating in a harsh environment. Failure would trigger faults in other areas. Failure would involve high repair costs
Weekly	Secondary equipment (not critical to continuous production). Difficult access for measurement
Monthly	Standard motors and similar equipment

sensor, known as 'pick-ups', can be used to suit the different parameters being monitored; these are reviewed below.

18.7.3.1 Absolute bearing vibration

This is generally measured on the outer loaded surface bearing housings, providing reliable criteria for establishing the condition of a machine fitted with ball or roller bearings. For the measurement of mechanical vibration and its conversion into an electrical signal, three different technologies are in common use:

(1) non-contacting pick-ups using the eddy current principle for the measurement of vibration displacement;
(2) electrodynamic pick-ups for the measurement of vibration velocity; and
(3) piezo-electric accelerometers for the measurement of vibration acceleration.

Damaged or worn rolling bearings will emit shock pulses which excite resonance in acceleration- or velocity-type pick-ups mounted on the loaded part of the bearing housing. The strength of individual pulses and the difference between strong and weak pulses provides the data for analysing the condition of the bearing. The choice of pick-up will be determined by the parameters to be monitored and the frequency ranges involved. The most common parameter in use is that of vibration velocity (or severity) which is typically measured in millimetres per second (V_{RMS}). At this point it must be stressed that adequate information regarding the 'acceptable' performance of a rolling bearing, or any other device being monitored, must be established and logged into the measurement database prior to any monitoring taking place.

Two standards have been published defining acceptable levels of vibration velocity for different types of rotating machinery commonly found in industry. These are:

BS 4675: Part 1: 1986 A Basis of Specifying Evaluation Standards for Rotating Machines with Operating Speeds of 10 to 200 Revolutions per Second
ISO 2372 Mechanical Vibration of Machines with Operating Speeds from 10 to 200 Revolutions/Second

Figure 18.18 is taken from ISO 2372 and shows the range of acceptable and unacceptable levels of vibration.

18.7.3.2 Relative shaft vibration

This is normally measured by using two non-contacting displacement transducers as pick-ups located at right angles to each other, in order to measure the vibration in two axes. Their principal application is in the measurement of the vibratory behaviour of a machine fitted with plain sleeve bearings.

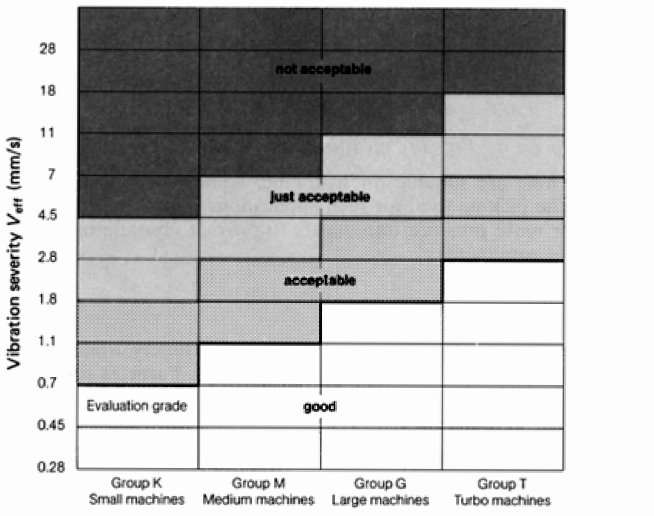


Figure 18.18 Evaluation limits for the vibrational behaviour of machines. (Courtesy of Schenk Ltd)

18.7.3.3 Axial shaft vibration

This term is used to describe the shift of a shaft in the axial direction. As with relative shaft vibration, measurements are normally made with non-contacting displacement transducers, sensing the position of the shaft end, step, collar or flange.

18.7.3.4 Eccentricity

Turbine rotors and similar shafts can bend due to gravity or the effects of temperature whilst at rest. This can be detected and measured during run-up by using non-contacting displacement pick-ups. Any deviation will be identified and initiate the shut down of the machine before any significant damage can occur.

18.7.3.5 Relative shaft expansion

Again this is an application for the non-contacting pick-up to measure any difference in the length of a shaft compared with its casing. Such differences frequently occur where, say, a turbine rotor will heat up more quickly than its housing.

18.7.3.6 Rotational speed

Several alternative technologies are available for measuring rotational speed and their individual use depends on the accuracy required, as well as the particular application. Pick-ups in common use are those of the non-contacting variety, both displacement and inductive proximity tranducers. Equally, a wide range of tachometers using different technologies is commercially available.

18.7.3.7 Temperature measurement

This may be used in a variety of situations where friction occurs, or where changes may not necessarily indicate a pending failure condition, for example, temperature measure-

ments of lubricants, motor or generator windings can provide valuable informative data on machine loadings at the time of measurement. In general, temperature measurements are made using thermocouples or resistance thermometers.

18.7.3.8 Pressure measurements

These are suitable for hydraulic, liquid or gas installations. The pick-up used for such applications usually comprises one or more pressure transducers to provide absolute or relative measurements.

18.7.3.9 Oil debris

This is another application for any machinery which has oil lubrication, hydraulic controls or drives. Particles suspended in oil may be isolated and identified as to their source. Debris may be collected using magnetic plugs or filters, from which it will be possible to identify debris in the range 100–1000 μm. For smaller particle sizes, ferrography enables microscopic examination and spectrographic analysis will determine the presence of particles of less than 10 μm in size.

One final note on condition monitoring concerns false alarms. Throughout the condition-monitoring process it is important that, where possible, duplicate sensors and circuits are used in order to highlight erroneous signals. If the application is suitable, then two different parameters should be measured before a failure signal is initiated. This obviously would not preclude investigation of signals from one pick-up which may be monitoring an apparently degrading situation. Such situations may already have been predicted as a result of trend data indicating an abnormal situation which may have been developing over a period of time.

18.8 Inventory and maintenance

18.8.1 Inventory objectives versus maintenance ideals

In some organisations, inventory may be the responsibility of the maintenance department, whilst in others it may be that of an independent function. More often than not the responsibility is a shared one, perhaps with maintenance being responsible for its own day-to-day requirements, drawing these from a centralised inventory and/or procurement department.

Whatever the organisation, the same fundamental conflict of interests will exist, whereby the maintenance department would like to have a quantity of every part instantly available and the inventory management (with an eye to the costs of holding the stock in readiness), will have other ideas. In reality, a balance has to be struck between availability of spare parts and stock holding costs. It is possible to do a complete cost analysis on this situation and, assuming the spares call-off rates are known, calculate the costs of holding each part of the equipment at each stocking echelon against the cost of delivery of the part from a higher echelon, together with the associated lost waiting time by the maintenance function. This, of course, does not take into account the lost production of the user, or his perception of the service that he is getting from the maintenance function. In practice, therefore, inventory must be optimised to provide the best possible service within whatever downtime is permissible and cost level is considered appropriate. This section is devoted to methods which can be adopted to achieve this objective.

Inventory is usually stocked at a number of different echelons, typically:

(1) with the equipment, perhaps in a locker only accessible to the maintenance engineer (unless there is an operator maintenance policy in being, which requires the operator to fit new parts or consumable supplies);
(2) with the maintenance engineer, in his/her vehicle if he/she is mobile, or again in a locker if he/she is stationed close to the equipment;
(3) in a local store, geographically central to a number of engineers; and
(4) in a central store, supplying parts direct to the base location of the engineer, or via a local store.

At each echelon, a goal should be set by which the efficiency of the supply can be measured. One method used comprises a logistics level of service (LOLOS) factor. A LOLOS of 98%, for example, would mean that a demand from stock can be met 98 times out of every 100 requests, the other two requests having to be referred for supplying from the next highest stocking echelon. It is important to understand that this factor applies to the overall operation of the store and not just the individual parts. For critical or high usage parts, sufficient stock must be held to meet almost every demand, otherwise a LOLOS of 98% overall would not be obtainable. Parts are normally classified according to their criticality and usage rates, providing a basis on which all inventory echelon and quantity decisions should be made. Remembering that the highly critical parts are those which affect safety and availability, these must therefore be readily available to the maintenance engineer, unless they have an extremely low replacement-rate prediction.

The factors which must be taken into account when determining inventory levels are:

(1) siting of stock at different locations;
(2) elapsed time betwen demand for replenishment or emergency order from one echelon to the next higher one;
(3) call-off by part number (replacement rate);
(4) stock holding cost;
(5) elapsed time between sending part for repair and its return as a fully serviceable part; and
(6) the LOLOS goal for each stock holding level.

Ideally, therefore, the overall inventory objective must be to minimise the direct and indirect costs of stock holding, whilst providing the highest LOLOS to the maintenance engineer. This can only be achieved through an awareness of all the cost implications, careful selection of parts, stocking echelons and order quantities.

18.8.2 Cost of stock holding

Stock is transferred in and out of the inventory system either at a standard cost or, if a profit-centre operation is in existence, a marked-up transfer price will apply to parts movements between some or all of the stocking echelons. The cost of holding stock depends on the following factors:

(1) interest which could be gained on the open market on capital which is invested in the stock;
(2) cost of storage facilities (rents, rates, insurance, etc.);
(3) cost of operation (staffing, transport and control systems); and
(4) deterioration and obsolescence.

An overhead figure of 25% of the cost of each part per annum would typically cover the costs of stock holding.

The 80/20 (Pareto) rule when applied to spare parts states that 20% of the parts account for 80% of the cost of parts usage (see *Figure 18.19*). Where these parts can be identified, this immediately indicates the priorities for close monitoring.

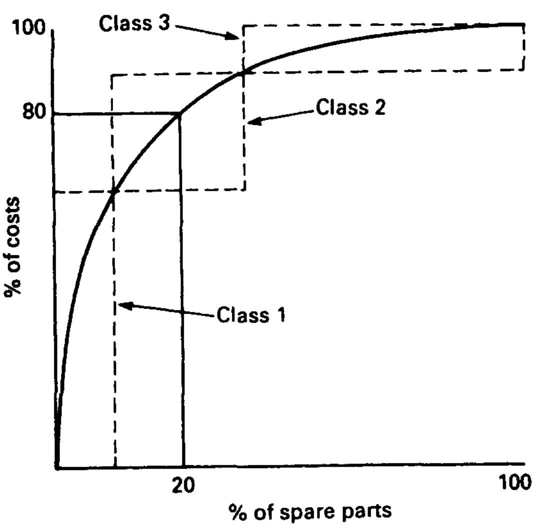


Figure 18.19 The 80/20 rule applied to spares

It is often convenient to classify these parts into three groups, as shown in *Figure 18.19*, allocating a usage value to each part number.

Usage value = Item cost × Usage per annum

From the point of view of controlling costs, priority should therefore be given to those items falling within class I and these should be subjected to strict inventory control.

18.8.2.1 Cost of stock-out

This brings into focus the implications of not having a part available when it is required. Specifically, these include:

(1) loss of equipment availability with consequential cost incurred in loss of output;
(2) additional costs involved in rushing small quantities of parts from the central warehouse to the maintenance engineer; and
(3) loss of goodwill, confidence in the equipment and the maintenance support, leading to loss of future sales when the time comes for equipment replacement.

The first two costs can obviously be quantified for any particular situation, but the third should not be overlooked, even if it is somewhat subjective, as this represents the future of the company (and the maintenance department!).

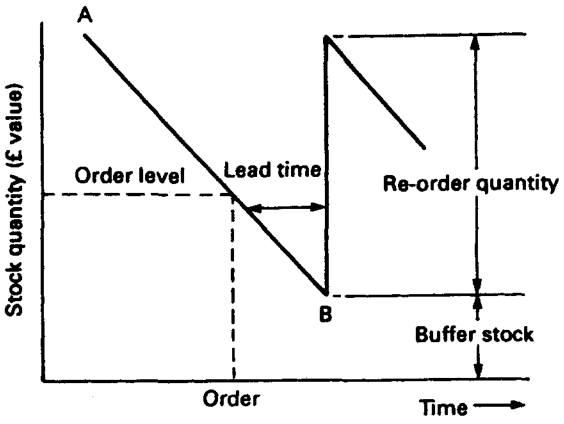


Figure 18.20 Stock movement and reorder trigger points

18.8.3 Safety stock and reorder quantities

However well planned the methods of supplying parts to the maintenance engineer, the system will only be successful if it is supported by effective stock-control procedures. In particular, keeping realistic levels of safety stock and ordering optimum quantities of new parts within the supplier's lead time. Normally, reorder systems depend on reaching a 'trigger' stock level at the central warehouse, or on the basis of a periodic review. The former is obviously the more efficient, and ideally lends itself to a computerised stock-control system. Where reordering is dependent on a periodic review, then the minimum stock levels must be sufficient to cover the call-off of parts between reviews, as well as during the reordering and replenishment cycle. Quite often a combination of both systems is used—the 'trigger' stock level for the more expensive parts and the periodic review for the cheaper parts.

Figure 18.20 shows diagrammatically the movement of stock against time. The slope of AB is based on historical usage, and

Table 18.12 Table to show the economic order quantity

Order quantity	*Orders per year*	*Ordering cost* (£)	*Average stock*	*Holding cost* (£)	*Total cost* (£)
500	20	1000	250	312	1250
1000	10	500	500	625	1125
1500	6.7	333	750	937	1270
2000	5	250	1000	1250	1500

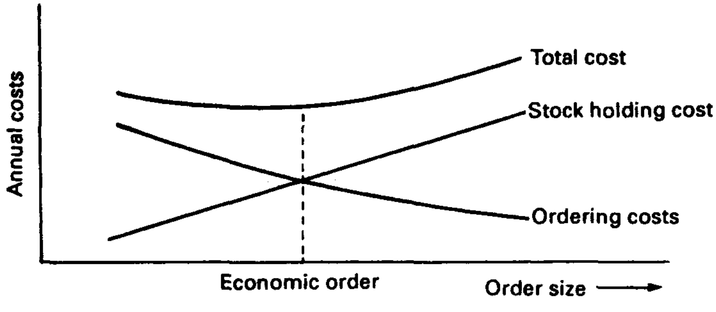


Figure 18.21 Determination of economic order quantity

the starting point in the parts-ordering activity is to establish the anticipated lead time involved between identifying the need to reorder and the receipt of the new stock. Having done this,

Buffer stock = Average usage during ordering lead time/3

or

Order level = 3 × Buffer stock

Note that the factor of one-third (or three), should be regarded as a good starting point, and experience will determine whether it is practical to reduce it.

The methodology involved in determining the economic order quantity (as shown in *Figure 18.21*) is best demonstrated by an example. Suppose that the annual spares usage of a particular part is 10 000, and its cost is £5. The inventory overheads are 25% of the parts in stock and the cost of raising an order and handling its delivery is £50. In *Table 18.12*,

$$\text{Orders per year} = \frac{10\,000}{\text{Order quantity}}$$

Ordering cost = Orders per year × 50

Average stock variation = Order quantity/2

Note that, as the level of the buffer stock depends on usage and ordering lead times, it remains constant in this calculation and may be ignored.

Holding cost = Average stock variation × (0.25 × 5)

Total cost = Stock holding cost + Ordering cost

As can be seen from *Table 18.12*, the economic ordering quantity would be around 1000. It should be added that, if a supplier offers a discount for a larger quantity, then this effect would have to be incorporated into the figures in *Table 18.12*.

There are several good software packages that embody the inventory principles described above, both as stand-alone systems and those integrated with maintenance management systems.

Applicable standards

BRITISH STANDARDS INSTITUTION, *BS 3843 Guide to terotechnology (the economic management of assets)*, Milton Keynes (1991)

BRITISH STANDARDS INSTITUTION, *BS 4675 Mechanical vibration in rotating machinery*, Milton Keynes (1986)

INTERNATIONAL ORGANISATION FOR STANDARDISATION, *ISO 2372 Mechanical vibration of machines with operating speeds of 10 to 200 revolutions per second*, Geneva

Further reading

FILE, W. T., *Cost Effective Maintenance—Design and Implementation*, Butterworth-Heinemann, Oxford (1991)

KELLY, A., *Maintenance Planning and Control*, Butterworth-Heinemann, Oxford (1984)

NEALE, M., HENRY, T. *et al.*, Condition based maintenance. *Maintenance*, **1**, No. 4. Conference Communication, Farnham, Surrey (1986)

NEIBEL, B. W., *Engineering Maintenance Management*, Marcel Dekker, New York (1985)

THABIT, S. S., Life cycle costing—a decade of progress. *Chartered Mechanical Engineer*, Bury St Edmunds, Suffolk (May 1983)

Table A17.2 Single sampling plans for normal inspection*

Acceptable quality levels (normal inspection), AQL 0.010–1.0 (each cell: *Ac Re*)

Sample size code letter	Sample size	0.010	0.015	0.025	0.040	0.065	0.10	0.15	0.25	0.40	0.65	1.0
A	2	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
B	3	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
C	5	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
D	8	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
E	13	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	0 1
F	20	↓	↓	↓	↓	↓	↓	↓	↓	↓	0 1	↑
G	32	↓	↓	↓	↓	↓	↓	↓	↓	0 1	↑	↓
H	50	↓	↓	↓	↓	↓	↓	↓	0 1	↑	↓	1 2
J	80	↓	↓	↓	↓	↓	↓	0 1	↑	↓	1 2	2 3
K	125	↓	↓	↓	↓	↓	0 1	↑	↓	1 2	2 3	3 4
L	200	↓	↓	↓	↓	0 1	↑	↓	1 2	2 3	3 4	5 6
M	315	↓	↓	↓	0 1	↑	↓	1 2	2 3	3 4	5 6	7 8
N	500	↓	↓	0 1	↑	↓	1 2	2 3	3 4	5 6	7 8	10 11
P	800	↓	0 1	↑	↓	1 2	2 3	3 4	5 6	7 8	10 11	14 15
Q	1 250	0 1	↑	↓	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22
R	2 000	↑	↑	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	↑

Acceptable quality levels (normal inspection), AQL 1.5–1000 (each cell: *Ac Re*)

Sample size code letter	Sample size	1.5	2.5	4.0	6.5	10	15	25	40	65	100	150	250	400	650	1000
A	2	↓	↓	↓	0 1	↓	↓	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	30 31
B	3	↓	↓	0 1	↑	↓	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	30 31	44 45
C	5	↓	0 1	↑	↓	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	30 31	44 45	↑
D	8	0 1	↑	↓	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	30 31	44 45	↑	↑
E	13	↑	↓	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	30 31	44 45	↑	↑	↑
F	20	↓	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	↑	↑	↑	↑	↑	↑
G	32	1 2	2 3	3 4	5 6	7 8	10 11	14 15	21 22	↑	↑	↑	↑	↑	↑	↑
H	50	2 3	3 4	5 6	7 8	10 11	14 15	21 22	↑	↑	↑	↑	↑	↑	↑	↑
J	80	3 4	5 6	7 8	10 11	14 15	21 22	↑	↑	↑	↑	↑	↑	↑	↑	↑
K	125	5 6	7 8	10 11	14 15	21 22	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
L	200	7 8	10 11	14 15	21 22	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
M	315	10 11	14 15	21 22	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
N	500	14 15	21 22	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
P	800	21 22	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
Q	1 250	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
R	2 000	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑

* From ISO 2859 1: 1989 (E).

↓. Use first sampling plan below arrow. If sample size equals, or exceeds, lot or batch size, carry out 100% inspection.

↑. Use first sampling plan above arrow.

Ac. Acceptance number.

Re. Rejection number.

Table A17.3 Double sampling plans for normal inspection†

Sample size code letter	Sample	Sample size	Cumulative sample size	Acceptable quality levels (normal inspection)										
				0.010	0.015	0.025	0.040	0.065	0.10	0.15	0.25	0.40	0.65	1.0
				Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re
A				↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
B	First	2	2	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
	Second	2	4	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
C	First	3	3	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
	Second	3	6	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
D	First	5	5	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
	Second	5	10	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
E	First	8	8	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	*
	Second	8	16	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	
F	First	13	13	↓	↓	↓	↓	↓	↓	↓	↓	↓	*	↑
	Second	13	26	↓	↓	↓	↓	↓	↓	↓	↓	↓		↑
G	First	20	20	↓	↓	↓	↓	↓	↓	↓	↓	*	↑	↓
	Second	20	40	↓	↓	↓	↓	↓	↓	↓	↓		↑	↓
H	First	32	32	↓	↓	↓	↓	↓	↓	↓	*	↑	↓	0 2
	Second	32	64	↓	↓	↓	↓	↓	↓	↓		↑	↓	1 2
J	First	50	50	↓	↓	↓	↓	↓	↓	*	↑	↓	0 2	0 3
	Second	50	100	↓	↓	↓	↓	↓	↓		↑	↓	1 2	3 4
K	First	80	80	↓	↓	↓	↓	↓	*	↑	↓	0 2	0 3	1 4
	Second	80	160	↓	↓	↓	↓	↓		↑	↓	1 2	3 4	4 5
L	First	125	125	↓	↓	↓	↓	*	↑	↓	0 2	0 3	1 4	2 5
	Second	125	250	↓	↓	↓	↓		↑	↓	1 2	3 4	4 5	6 7
M	First	200	200	↓	↓	↓	*	↑	↓	0 2	0 3	1 4	2 5	3 7
	Second	200	400	↓	↓	↓		↑	↓	1 2	3 4	4 5	6 7	8 9
N	First	315	315	↓	↓	*	↑	↓	0 2	0 3	1 4	2 5	3 7	5 9
	Second	315	630	↓	↓		↑	↓	1 2	3 4	4 5	6 7	8 9	12 13
P	First	500	500	↓	*	↑	↓	0 2	0 3	1 4	2 5	3 7	5 9	7 11
	Second	500	1 000	↓		↑	↓	1 2	3 4	4 5	6 7	8 9	12 13	18 19
Q	First	800	800	*	↑	↓	0 2	0 3	1 4	2 5	3 7	5 9	7 11	11 16
	Second	800	1 600		↑	↓	1 2	3 4	4 5	6 7	8 9	12 13	18 19	26 27
R	First	1 250	1 250	↑	↑	0 2	0 3	1 4	2 5	3 7	5 9	7 11	11 16	↑
	Second	1 250	2 500	↑	↑	1 2	3 4	4 5	6 7	8 9	12 13	18 19	26 27	↑

Sample size code letter	Sample	Acceptable quality levels (normal inspection)														
		1.5	2.5	4.0	6.5	10	15	25	40	65	100	150	250	400	650	1000
		Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re	Ac Re
A		↓	↓	↓	*	↓	↓	*	*	*	*	*	*	*	*	*
B	First	↓	↓	*	↑	↓	0 2	0 3	1 4	2 5	3 7	5 9	7 11	11 16	17 22	25 31
	Second	↓	↓		↑	↓	1 2	3 4	4 5	6 7	8 9	12 13	18 19	26 27	37 38	56 57
C	First	↓	*	↑	↓	0 2	0 3	1 4	2 5	3 7	5 9	7 11	11 16	17 22	25 31	↑
	Second	↓		↑	↓	1 2	3 4	4 5	6 7	8 9	12 13	18 19	26 27	37 38	56 57	↑
D	First	*	↑	↓	0 2	0 3	1 4	2 5	3 7	5 9	7 11	11 16	17 22	25 31	↑	↑
	Second		↑	↓	1 2	3 4	4 5	6 7	8 9	12 13	18 19	26 27	37 38	56 57	↑	↑
E	First	↑	↓	0 2	0 3	1 4	2 5	3 7	5 9	7 11	11 16	17 22	25 31	↑	↑	↑
	Second	↑	↓	1 2	3 4	4 5	6 7	8 9	12 13	18 19	26 27	37 38	56 57	↑	↑	↑
F	First	↓	0 2	0 3	1 4	2 5	3 7	5 9	7 11	11 16	↑	↑	↑	↑	↑	↑
	Second	↓	1 2	3 4	4 5	6 7	8 9	12 13	18 19	26 27	↑	↑	↑	↑	↑	↑
G	First	0 2	0 3	1 4	2 5	3 7	5 9	7 11	11 16	↑	↑	↑	↑	↑	↑	↑
	Second	1 2	3 4	4 5	6 7	8 9	12 13	18 19	26 27	↑	↑	↑	↑	↑	↑	↑
H	First	0 3	1 4	2 5	3 7	5 9	7 11	11 16	↑	↑	↑	↑	↑	↑	↑	↑
	Second	3 4	4 5	6 7	8 9	12 13	18 19	26 27	↑	↑	↑	↑	↑	↑	↑	↑
J	First	1 4	2 5	3 7	5 9	7 11	11 16	↑	↑	↑	↑	↑	↑	↑	↑	↑
	Second	4 5	6 7	8 9	12 13	18 19	26 27	↑	↑	↑	↑	↑	↑	↑	↑	↑
K	First	2 5	3 7	5 9	7 11	11 16	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
	Second	6 7	8 9	12 13	18 19	26 27	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
L	First	3 7	5 9	7 11	11 16	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
	Second	8 9	12 13	18 19	26 27	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
M	First	5 9	7 11	11 16	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
	Second	12 13	18 19	26 27	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
N	First	7 11	11 16	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
	Second	18 19	26 27	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
P	First	11 16	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
	Second	26 27	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
Q	First	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
	Second	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
R	First	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
	Second	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑

↓. Use first sampling plan below arrow. If sample size equals, or exceeds, lot or batch size, carry out 100% inspection.

↑. Use first sampling plan above arrow.

Ac. Acceptance number.

Re. Rejection number.

* Use corresponding single sampling plan (or, alternatively, use double sampling plan below, where available).

† From ISO 2859 1: 1989 (E).

Table A17.4 Poisson cumulative probability tables (×1000) the probability of *Ac* or less defectives, where *np* is the average number of expected defectives

np	*Ac* 0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
0.02	980	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
0.04	961	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
0.06	942	998	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
0.08	923	997	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
0.10	905	995	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
0.15	861	990	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
0.20	819	982	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
0.25	779	974	998	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
0.30	741	963	996	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
0.35	705	951	994	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
0.40	670	938	992	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
0.45	638	925	989	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
0.50	607	910	986	998	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
0.55	577	894	982	998	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
0.60	549	878	977	997	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
0.65	522	861	972	996	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
0.70	497	844	966	994	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
0.75	472	827	959	993	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
0.80	449	809	953	991	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
0.85	427	791	945	989	998	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
0.90	407	772	937	987	998	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
0.95	387	754	929	984	997	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
1.00	368	736	920	981	996	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
1.10	333	699	900	974	995	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
1.20	301	663	879	966	992	998	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
1.30	273	627	857	957	989	998	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
1.40	247	592	833	946	986	997	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
1.50	223	558	809	934	981	996	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
1.60	202	525	783	921	976	994	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
1.70	183	493	757	907	970	992	998	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
1.80	165	463	731	891	964	990	997	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
1.90	150	434	704	875	956	987	997	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
2.00	135	406	677	857	947	983	995	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
2.20	111	355	623	819	928	975	993	998	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
2.40	91	308	570	779	904	964	988	997	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
2.60	74	267	518	736	877	951	983	995	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
2.80	61	231	469	692	848	935	976	992	998	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
3.00	50	199	423	647	815	916	966	988	996	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
3.20	41	171	380	603	781	895	955	983	994	998	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
3.40	33	147	340	558	744	871	942	977	992	997	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
3.60	27	126	303	515	706	844	927	969	988	996	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
3.80	22	107	269	473	668	816	909	960	984	994	998	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
4.00	18	92	238	433	629	785	889	949	979	992	997	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
4.20	15	78	210	395	590	753	867	936	972	989	996	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
4.40	12	66	185	359	551	720	844	921	964	985	994	998	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
4.60	10	56	163	326	513	686	818	905	955	980	992	997	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
4.80	8	48	143	294	476	651	791	887	944	975	990	996	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
5.00	7	40	125	265	440	616	762	867	932	968	986	995	998	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
5.20	6	34	109	238	406	581	732	845	918	960	982	993	997	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
5.40	5	29	95	213	373	546	702	822	903	951	977	990	996	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
5.60	4	24	82	191	342	512	670	797	886	941	972	988	995	998	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
5.80	3	21	72	170	313	478	638	771	867	929	965	984	993	997	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
6.00	2	17	62	151	285	446	606	744	847	916	957	980	991	996	999	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
6.20	2	15	54	134	259	414	574	716	826	902	949	975	989	995	998	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
6.40	2	12	46	119	235	384	542	687	803	886	939	969	986	994	997	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
6.60	1	10	40	105	213	355	511	658	780	869	927	963	982	992	997	999	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
6.80	1	9	34	93	192	327	480	628	755	850	915	955	978	990	996	998	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
7.00	1	7	30	82	173	301	450	599	729	830	901	947	973	987	994	998	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
7.20	1	6	25	72	156	276	420	569	703	810	887	937	967	984	993	997	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
7.40	1	5	22	63	140	253	392	539	676	788	871	926	961	980	991	996	998	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
7.60	1	4	19	55	125	231	365	510	648	765	854	915	954	976	989	995	998	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
7.80	0	4	16	48	112	210	338	481	620	741	835	902	945	971	986	993	997	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
8.00	0	3	14	42	100	191	313	453	593	717	816	888	936	966	983	992	996	998	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
8.50	0	2	9	30	74	150	256	386	523	653	763	849	909	949	973	986	993	997	999	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
9.00	0	1	6	21	55	116	207	324	456	587	706	803	876	926	959	978	989	995	998	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
9.50	0	1	4	15	40	89	165	269	392	522	645	752	836	898	940	967	982	991	996	998	999	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
10.00	0	0	3	10	29	67	130	220	333	458	583	697	792	864	917	951	973	986	993	997	998	999	1000	1000	1000	1000	1000	1000	1000	1000	1000
10.50	0	0	2	7	21	50	102	179	279	397	521	639	742	825	888	932	960	978	988	994	997	999	999	1000	1000	1000	1000	1000	1000	1000	1000
11.00	0	0	1	5	15	38	79	143	232	341	460	579	689	781	854	907	944	968	982	991	995	998	999	1000	1000	1000	1000	1000	1000	1000	1000
11.50	0	0	1	3	11	28	60	114	191	289	402	520	633	733	815	878	924	954	974	986	992	996	998	999	1000	1000	1000	1000	1000	1000	1000
12.00	0	0	1	2	8	20	46	90	155	242	347	462	576	682	772	844	899	937	963	979	988	994	997	999	999	1000	1000	1000	1000	1000	1000
12.50	0	0	0	2	5	15	35	70	125	201	297	406	519	628	725	806	869	916	948	969	983	991	995	998	999	999	1000	1000	1000	1000	1000
13.00	0	0	0	1	4	11	26	54	100	166	252	353	463	573	675	764	835	890	930	957	975	986	992	996	998	999	1000	1000	1000	1000	1000
13.50	0	0	0	1	3	8	19	41	79	135	211	304	409	518	623	718	798	861	908	942	965	980	989	994	997	998	999	1000	1000	1000	1000
14.00	0	0	0	0	2	6	14	32	62	109	176	260	358	464	570	669	756	827	883	923	952	971	983	991	995	997	999	999	1000	1000	1000
14.50	0	0	0	0	1	4	10	24	48	88	145	220	311	413	518	619	711	790	853	901	936	960	976	986	992	996	998	999	999	1000	1000
15.00	0	0	0	0	1	3	8	18	37	70	118	185	268	363	466	568	664	749	819	875	917	947	967	981	989	994	997	998	999	1000	1000
16.00	0	0	0	0	0	1	4	10	22	43	77	127	193	275	368	467	566	659	742	812	868	911	942	963	978	987	993	996	998	999	999
17.00	0	0	0	0	0	1	2	5	13	26	49	85	135	201	281	371	468	564	655	736	805	861	905	937	959	975	985	991	995	997	999
18.00	0	0	0	0	0	0	1	3	7	15	30	55	92	143	208	287	375	469	562	651	731	799	855	899	932	955	972	983	990	994	997
19.00	0	0	0	0	0	0	1	2	4	9	18	35	61	98	150	215	292	378	469	561	647	725	793	849	893	927	951	969	980	988	993
20.00	0	0	0	0	0	0	0	1	2	5	11	21	39	66	105	157	221	297	381	470	559	644	721	787	843	888	922	948	966	978	987
22.00	0	0	0	0	0	0	0	0	1	2	4	8	15	28	48	77	117	169	232	306	387	472	556	637	712	777	832	877	913	940	959
24.00	0	0	0	0	0	0	0	0	0	0	1	3	5	11	20	34	56	87	128	180	243	314	392	473	554	632	704	768	823	868	904
26.00	0	0	0	0	0	0	0	0	0	0	0	1	2	4	8	14	25	41	65	97	139	190	252	321	396	474	552	627	697	759	813
28.00	0	0	0	0	0	0	0	0	0	0	0	0	1	1	3	5	10	18	30	48	73	106	148	200	260	327	400	475	550	623	690
30.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	4	7	13	22	35	54	81	115	157	208	267	333	403	476	548

19 Ergonomics

E N Corlett

Contents

19.1 Introduction 19/3
19.2 Robotics 19/4
19.3 Computer-aided manufacture 19/5
19.4 The automated office 19/5
19.5 Manual assembly 19/6
19.6 The economics of ergonomics 19/6
19.7 Conclusions 19/7

19.1 Introduction

To argue for the importance of manufacturing in the nation's economy is probably unnecessary for readers of this book. It is an activity which we must pursue efficiently for the benefit of society at large; it is the vehicle by which the conditions of society are improved and the contributions of many aspects of intellectual activity are made of general utility. However, it can have its negative aspects, as we all know. The conditions of working people, and the ruination of nature, as a result of industrial working practices are the stuff of industrial history, of many social studies and, unfortunately, the results of much industrial activity today.

The enormous scale of environmental destruction and the widening of the realisation of the exploitation of people in some manufacturing situations world-wide, have brought a recognition that we must introduce new concepts into the assessment of industrial efficiency. Manufacturing is not the key factor in the economy before which all others must be subordinated, but is one of those which contributes to an acceptable form of society. Society comprises both the people who buy the manufactured products and the people who make them. To put it at its lowest, it would be foolish to disable (in any way) those working in manufacturing and to reduce their likelihood of buying the products. On a slightly higher level, the costs of making the product, in either economic, social or ecological terms, must not outweigh the benefits.

This preliminary is necessary because when we talk of occupational ergonomics we are dealing with the interactions between people and their work. As is shown in more detail later, this covers the physical, environmental, cognitive, organisational and social interactions to which we are all susceptible. It is not enough, for example, to consider just the reach distances and movement times in the design of a workplace, because optimum human performance is not achieved just by minimising these two factors. For a machine they may be appropriate, although the good designer would want to consider wear and expected machine life, maintenance and repairability, cost, manufacturability and several other factors. Therefore, even for a machine this pair of criteria is not enough, yet how many people believe that these criteria are adequate for judging the fitness of a person's workplace?

Underlying such a view is the perspective on a person as a simple machine, whose operation and motivation are independent of the effects of work activities. This implies a relatively simple system influenced by only a few variables of the most obvious kind. To reduce the argument to these terms is to reveal its absurdity. It also suggests that our personal knowledge of the complexity of human relationships and the effects of knowledge, difficulties and experiences on the performance of work as we all know of them, somehow does not apply to people doing industrial work. *They* are different, and can be categorised and valued broadly in terms of work study variables alone.

As a minimum we suggest that, when choosing the criteria on which to assess whether a work station is adequate, we must look beyond output in the short term; the interest is in quality performance in the long term. This requires that the work can be done day after day without detriment to the person's physical or mental health, with acceptable levels of accuracy and correctness, and with motivation. If any of these factors are absent, the firm will lose money and the workers are wasting their time. But this, i.e. that it shall not be harmful, is a negative level of acceptability. Why should acceptability not be set in positive terms?

Many professional people use skills in their work which they value, and take pleasure in their jobs. They put more into their work than is formally expected and enjoy the respect and good name which their performance engenders. Success, in such cases, breeds further success. Why should this situation not prevail at the shop floor or general office level? Why not set job-quality criteria beyond the minimal ('animal') level, in order to recognise and cater for the human dimensions of people?

Some readers may, by now, be thinking that such discussions go beyond ergonomics. A commonly held view, when one sees such phrases as 'ergonomically designed seating', is that the subject is primarily concerned with human body dimensions and their relationships with devices. But a more reliable short definition is 'matching work and its environment to people', and the environment can be illustrated as shown in *Figure 19.1*.

In the centre of the figure there is a 'man–machine interface'. Some process is influenced by a person, who senses the changes to the process, giving feedback of the effects on the process. The person is presented as a physiological system enclosing the sensing and effecting organs (the latter, of course, including speech), which in turn are linked by the cognitive system. Within this is an emotional aspect.

A feature of this diagram is that, wherever a boundary exists between any two areas, there are interactions across it. That is to say, taking as an example the sensory organs and the physical environment, the ambient lighting conditions will influence the ability to see, just as the ambient noise levels will affect what the person can hear. So the diagram warns us to explore some possible interaction effects which will influence a person's ability to achieve their objectives.

It is not possible in a short chapter to discuss all the areas for interaction, but in the light of the above discussion, it can be seen that the work organisation (e.g. piecework, shiftwork or group working) will affect people's emotional responses, as will their experiences and expectations arising from their social life (e.g. what their friends do, or what they see on television). These emotional responses will, in turn, influence the decisions they make concerning their work activities, and can directly affect their immediate responses to work demands. An example of the latter is the less well controlled driving activity that follows from seeing a serious accident or being angered by the behaviour of a fellow motorist.

Before describing some particular areas where ergonomics and advanced manufacturing technologies meet, some comment on the links between ergonomics and work study are appropriate. These two contributions to productivity are not totally opposed to one another, indeed there is no substitute, for example, to the use of predetermined motion–time systems in the evaluation of output and manning levels for new production systems involving people. There are also many useful measuring and recording procedures for dealing with work activities. The new contributions which ergonomics is making to work study are in the better understanding of the relationships between the arrangements which work study proposes, and in the performance, health and fatigue of people implementing these proposals. As mentioned earlier, the somewhat mechanistic view of how people do their work, presented by the pioneers at the turn of the century, is giving way to more human-oriented concepts and requirements. The principles of workplace design[1] given in *Table 19.1* are an example, replacing the limited concepts of early work study 'principles of easy movement', or similar titles. They offer an ordered sequence of requirements which match more closely the physiological needs of the human body. Increased knowledge of fatigue and recovery, as well as a better understanding of what people expect from work, enable jobs to be more productive with less of the aggravations which have been seen as a traditional part of, at least, much of British working relationships.

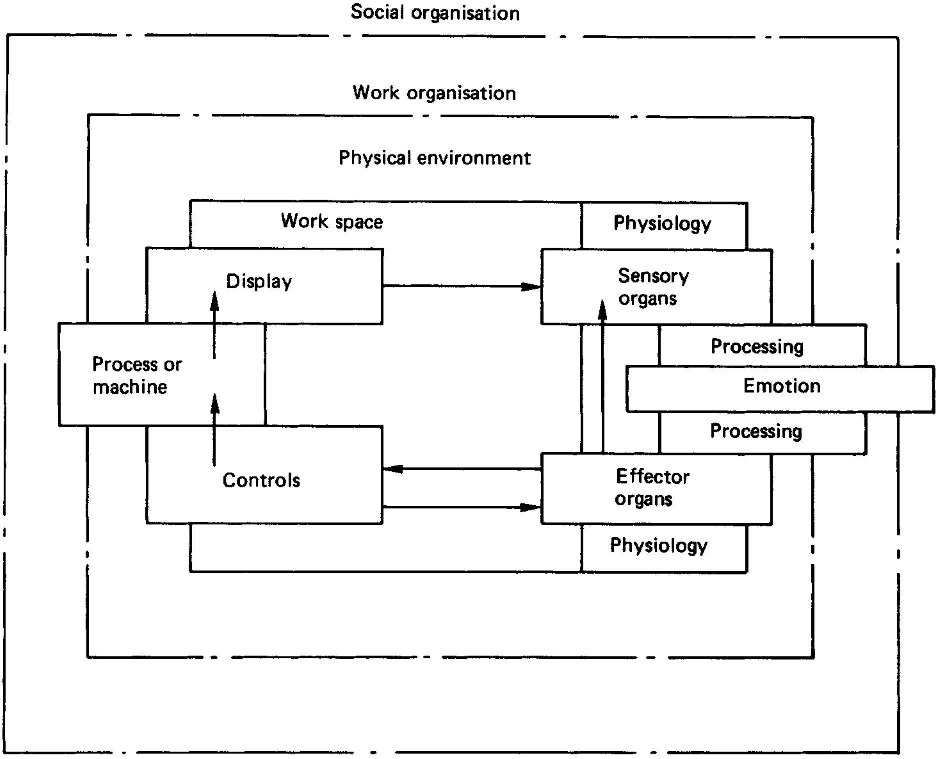


Figure 19.1 The man–machine system within its interactive environments

To close this section, one point which is important to ergonomics must be emphasised. The point of view of the ergonomist is that of the person doing the work and not as an observer watching the work being done. The sort of questions asked are 'How does this job match the person's capacities?' and 'What factors in the work situation give rise to these errors?' The direction of effort is to understand how the situation affects a person's performance, and hence how to change it so that people can achieve their objectives. It is *not* to assume that the workplace is 'optimum' and the person must be changed. Changing the circumstances enables people to modify their activities to give good performance. There are, of course, situations where we want people to change, hence the need for training, for example, but removing the obstructions to good performance is the initially preferred direction of the ergonomist.

19.2 Robotics

In the UK it is required that people and robots are separated, and access to the latter requires their isolation from the power supply or a reduction in their speed of movement. Until further developments in hardware and software this probably remains the safest procedure, for people and robots do not mix well. People cannot repeat the same actions consistently throughout a working day, nor can they continually bear in mind the actions of a neighbouring robot, to always be sure that they are not in its way. Although multibackup avoidance systems for robots are possible, they are expensive and rarely used; to assume the total reliability of a single level of control system for the robot and its work area is not acceptable.

Many ergonomics investigations have concentrated on the control box/teach control. The control of a robot from a separate control box has something in common with controlling the remote handling systems used in nuclear fuel processing plants. The robot, as it moves, changes its orientation so that what was at the bottom might now be at the top, or moving it to the left when its gripper is away from you, might require moving the control to the right when it is towards you. Even very experienced operators make frequent errors using these devices, and these are increased if they have two or three different robots, each with its individual design of teach control.

People's perceptions of movement do not always match reality. Nagamachi[2] has demonstrated that if people are asked to estimate whether they have enough time to pick up an object underneath a moving robot arm, they always misjudge its movement speed if it is moving towards them, believing they have more time than there really is. Judgement of clearances is also relatively poor by people, i.e. they will try to get through spaces which are too small. The use of computer graphics to assess pinch points in a computer's work space has been described by Yong *et al.*[3] and clearance problems are readily assessed by means of this technique.

Table 19.1 Ergonomically appropriate principles for use in the design of work and work spaces

1. The worker should be able to maintain an upright and forward facing posture during work
2. Where vision is a requirement of the task, the necessary work points must be adequately visible with the head and trunk upright or with just the head inclined slightly forward
3. All work activities should permit the worker to adopt several different, but equally healthy and safe, postures without reducing capability to do the work
4. Work should be arranged so that it may be done, at the worker's choice, in either a seated or standing position. When seated, the worker should be able to use the backrest of the chair at will, without necessitating a change of movements
5. The weight of the body, when standing, should be carried equally on both feet, and foot pedals designed accordingly
6. Work activities should be performed with the joints at about the midpoint of their range of movement. This applies particularly to the head, trunk and upper limbs
7. Where muscular force has to be exerted it should be by the largest appropriate muscle groups available and in a direction colinear with the limbs concerned
8. Work should not be performed consistently at or above the level of the heart; even the occasional performance where force is exerted above heart level should be avoided. Where light hand work must be performed above heart level, rests for the upper arms are a requirement
9. Where a force has to be exerted repeatedly, it should be possible to exert it with either of the arms, or either of the legs, without adjustment to the equipment
10. Rest pauses should allow for all loads experienced at work, including environmental and information loads, and the length of the work period between successive rest periods

19.3 Computer-aided manufacture

When machines were operated by hand, operators understood the whole machine system, how their actions affected surface finish, shape and output and how the machine's responses influenced the component. Whilst machines have become more integrated and their processing more complex, their operators have become more remote, in the sense that their control over what happens is through software rather than via direct influence. To achieve control now requires a 'mental model' of what the system does. An analogy is our knowledge of our local town. If our most direct way is blocked, during a journey, by a road being closed, we can 'reprogramme' our trip using side alleys and back streets to minimise the distance we have to travel. In the case of the operators of complex systems, they must understand the system sufficiently to know how it responds to their attempts to control it, under any eventuality. It is not correct to assume that this is always best achieved by requiring the operator to follow the technical structure precisely. The form of presentation of information about the state of the process can assist or hinder the operator, whilst the way in which the program is structured can be crucial.

The difference can be exemplified by comparing a geographical representation of a railway system with the formalised display seen in a city underground map or a modern railway system control centre. The displays in these two situations are formalised to present the general directions and relationships between stations and other features relevant to the traveller or controller. To underground travellers the fact that their westbound train may at some time be travelling north is of no concern, because their mental map of the city tells them that their destination is broadly to the west of their starting point.

This is a broad subject, much studied by systems analysts and those who design process plant and nuclear generating plant control rooms. Not only does it have a major impact on the ease of learning and understanding the operation of complex manufacturing systems, but good design in this area can increase safety and reduce errors. It is obvious that, in emergencies, the controls, information presentation, etc., must be highly compatible with the operators' perceptions of the behaviour of the process. This compatibility is also important when the operator is tired, is on shift work or is interrupted or carries additional load due to some unforeseen circumstance. To put very expensive plant at risk by skimping on this key interface area is to create the sort of situation which gives rise to many dramatic tragedies, including the Three Mile Island atomic power station failure.

19.4 The automated office

Computation in the office has produced a revolution in operating practices. In many firms, the office and factory floor are tending to merge to bring information and control closer to the point of manufacture. Extensive studies of the appropriate arrangements of keyboards, screen layouts, lighting, seat and table designs have been made (see, for example, Cakir *et al.*[4]). A major effect arising from introducing computers is the increased amount of work which is done at the keyboard. Typescripts, filing and message despatch to other users' computers reduce the need for people to move from their chairs, increase the time they spend in a limited range of postures, and permit an increased number of keystrokes per day.

An extension of Tayloristic practices has led some companies to monitor keystrokes per day as a performance indicator, whilst others try pacing the keyboard operator by presenting information at a rate not under the operator's control. Several studies may be marshalled against these practices; perhaps the first point to be made is that the important criterion is most likely to be accuracy rather than speed, the old adage 'garbage in, garbage out' still holds, with the added comment that the garbage must be paid for. McAtamney[5] has demonstrated that pacing skilled keyboard operators even at a speed equivalent to their freely chosen pace, increases their error rate enormously, and that this does not appear to change with practice. What she also demonstrated was a recognisable increase in activity in the trapezius muscles, implying an increase in tension during the paced work periods.

This is interpreted as a contributory factor to the widespread effects (sometimes called 'repetitive strain injury') which arise in a considerable number of people due to too long periods at the keyboard. The constant tension in shoulder, neck and back muscles to maintain a suitable orientation at the

keyboard gives rise to aches in these body parts which, if continued, can become chronic. In addition, the repeated use of the fingers on the keys, where 100 000 strokes a day are possible, and 50 000 not unusual, gives rise to, *inter alia*, inflammation of the tendon sheath in the wrist, causing pain and loss of grip strength. This damage, if the activity is persisted in, may result in permanent injury; certainly the initial symptoms should be heeded and the job changed to reduce exposure.[6]

As with so much in ergonomics, the key factor is variety of activity and posture (see *Table 19.1*). The temptation to long periods of sitting due to the flexibility that the computer provides can lead to neck, shoulder and back problems, and to eye strain because of the limited range of focusing of the eyes (and sometimes to the poor lighting arrangements provided). Another problem of long periods of sitting is the pooling of blood in the legs. Walking to perform various office tasks reduces this tendency, but 2 or 3 h of continuous sitting can result in significant swelling.[7] These changes are not such as the body normally experiences, which leads researchers to anticipate adverse consequences if measures are not taken to introduce more leg activity into the working day.

19.5 Manual assembly

The development of modern electronics has introduced new problems in their manufacture. Electrical products, in common with many others, are relatively large and their assembly requires many different movements and changes in posture. If cycle times are short the variety of postures is reduced and their frequency of use increased, giving rise to musculoskeletal symptoms due to over-use of a few muscles and joints. With electronic products, where they are assembled by hand, it is not unusual to have a 20-s cycle time during which the operator assembles eight or so components.

All of the assembly techniques require a pinch grip to hold, followed by an arching of the wrist to place the component perpendicularly into the printed circuit board. Simple arithmetic shows that this will involve over 10 000 cycles a day of the same activity, and wrist problems do occur in assembly-line workers. Strictly speaking, this is not a job for a person at all, because the joints, muscles, etc., are not designed for such extensive use of a limited number of actions (to say nothing of the limited mental activity required). As an interim measure, reducing wrist flexure, increasing the variety of movement by careful board design, and job rotation or enlargement to include some inspection or other tasks, would probably eliminate the problems completely. Whether or not the company sees this as economical is the concern discussed in Section 19.6.

The mass production of advanced electronic products is not always as modern as the products being produced. Short cycle times are still seen as the most productive method to adopt, with paced lines in widespread use. Surprisingly, the physical damage which a proportion of operators experience due to these highly repetitive conditions is not fully recognised by either management or some medical personnel, whilst the great majority of those who create these workplaces are entirely ignorant of the consequences of their designs.

The wider consequences in terms of work attitudes, motivation and quality performance are also well documented, but even less well known to many whose job it is to be up to date in the creation of efficient manufacturing systems. Even much of the education, in the UK at least, provides very little modern knowledge in this area, being still locked into pre-War work study, wage incentive schemes and personnel relations laws and customs. Health-and-safety legislation, stimulated by Common Market pressure, is bringing changes, but there is still a long way to go if we are to match some of our major competitors' working performance and conditions.

19.6 The economics of ergonomics

It is not unusual to hear a manager discuss ergonomic proposals for workplace changes using a remark like, 'we can't afford to pay just to make the workers comfortable'. The implications are that the investment would not contribute (sufficiently) to productivity. Yet a moments thought would cause the manager to recognise that his/her secretary, the availability of tea and coffee, good desk and chair and easy chairs in the corner contributed much to comfort, and also much to the efficient and profitable use of the manager's time. Doing filing, searching for addresses or fending off unnecessary salesmen is not a good use of the time of a well-paid manager; equally, the appearance of an office can give confidence to a prospective customer, who might be unlikely to bring business to a manager working on a packing case in a corner of the stores. Exaggerated, perhaps, but the manager's comfort is a commercial advantage to the company—within limits of course!

What about our shop-floor workers? They might have worked for the company for several years, know its ways and its products. For them to spend part of their time and energies in overcoming needless obstacles of their equipment or needless bodily or mental fatigue is to be blind to the economic good of the company, which is paying them to produce its products, not to get tired and frustrated.

Many of our major competitors in Japan, Sweden or Germany, for example, do not make this mistake. In Japan, in particular, the people are a resource, not a cost, to be trained and retrained, encouraged and developed as part of the operational success of the enterprise. The costs of this activity are more than repaid by large quantities of high quality on-time products, a flexible workforce and people who, over the years, grow into a well-trained workforce with a considerable knowledge of the business.

This looks a costly operation to the typical Western manager. But how many companies relate their production policies to the costs of absence, lateness, hiring and training costs, and scrap, rework and returned deliveries?

Research in this field is rare, but one study[8] has examined the costs and returns from changes in the workplaces of women assembling telephone equipment. The data dealt with the situation before the ergonomic changes to the workplace (1967–1974) and the effects after the changes (1975–1987). Clearly such extensive data relate to no 'Hawthorne effect'. These authors assessed the costs arising from the purchase of ergonomic workstations, changes in ventilation and lighting and increased lighting costs. The benefits assessed were reductions in recruitment, training and instructors' salary costs and reductions in sick payments. All these factors were assessed for each of the years of the study. The costs to the company were NKr 338 992, whereas the savings were NKr 3 226 194; both figures reduced to a common value.

In this study no value was put on improvements in working relationships, which are reported in the paper, nor were changes in productivity evaluated and included. The documented reductions in sick leave and musculoskeletal ailments was a benefit to the workers themselves. The costs of such ailments in terms of a degraded social life are never borne by companies, but by society as a whole. The sharpening of health-and-safety legislation around the world may make a

change in this respect, it could be an unwise company which ignored the possibility of a penalty for having work conditions which made some of its workers ill or unemployable.

19.7 Conclusion

This chapter has brushed the surface of the complex relationships between people and their work. Although there are many publications and much research to draw on, it will, we hope, also be clear that, by adopting a changed perspective on people at work, there are many possibilities for change which do not need much science, but the application of the logical consequences of that change. If people have to reach and bend unduly, repeat simple mindless activities endlessly, or strain their eyes or ears to get the information needed for their job, then at the very least you can assume you are paying a salary for someone to overcome the equipment's problems, and not for making the product.

But this is not to say that it is all common sense. Compared with a person, an internal combustion engine is a relatively simple device, yet deep knowledge and experience are required to design a better one. Our very familiarity is liable to cause us to overlook the complexity: 'I am a person, surely what I feel should be done must be right'. Everyone, however, is different, and dealing with these variations is one of the skills of the ergonomically trained person. Understanding the variability of people also helps in understanding what the relevant factors are which can be altered, and what the short- and long-term changes in the result are likely to be.

In manufacturing engineering terms, recognising the contributions of people and designing to enable their full benefits to be obtained makes sense. Assuming all is technology and that people will fit in around it not only does not make sense, but has been demonstrated decade after decade not to work by companies who have fallen further and further behind in productivity terms. Modern industry needs a modern approach to its aim of achieving a competitive market position, and one of the major components of that approach must be arranging for its workforce, at all levels, to be in the best position to put all its efforts into achieving the company objectives.

References

1 CORLETT, E. N., The human body at work: new principles for designing workspaces and methods. *Management Services*, **22**(5), 20–52 (1978)
2 NAGAMACHI, M., Human factors of industrial robots and robot safety management in Japan. *Applied Ergonomics*, **17**(1), 9–18 (1986)
3 YONG, Y. F., TAYLOR, N. K. and BONNEY, M. C., CAD—an aid to robot safety. In *Robot Safety* (eds M. C. Bonney and Y. F. Yong), IFS Publications, Bedford (1985)
4 CAKIR, A., HART, D. J. and STEWART, T. F. M., *Visual Display Terminals*, Wiley, Chichester (1979)
5 MCATAMNEY, L., Some effects of task loading in a VDU-based task. In *Contemporary Ergonomics 1989* (ed. E. D. Megaw), Taylor & Francis, London (1989)
6 KNAVE, B. and WIDEBACK, P. G., *Work with Display Units*. Elsevier, Amsterdam (1987)
7 WINKEL, J., On foot swelling during prolonged sedentary work, and the significance of leg activity. *Arbete och Halsa 1985–35*. National Board of Occupational Safety and Health, Solna, Sweden (1985)
8 SPILLING, S., EITRHEIM, J. and AARAS, A., Cost benefit analysis of work environment investment at STK's telephone plant at Kongsvinger. In: *The Ergonomics of Working Postures* (eds. N. Corlett, J. Wilson and I. Manenica). Taylor & Francis, London (1986)

Index

Compiled by D.C. Tyler

Abrasives,
  see also Grinding/grinding wheels
  aluminium oxide, 8/4, 8/30
  belts,
    horizontal grinding, 8/29–30
    joints, 8/28
    performance, 8/28–9
    vertical grinding, 8/30
  blasting, 8/34
  coated,
    backstand machines, 8/29
    bonds, 8/27
    construction, 8/25–6
    contact wheels, 8/28
    cutting efficiency, 8/26
    cutting mechanism, 8/26–7
    flat bed finishes, 8/29
    grit types and sizes, 8/27
    storage and handling, 8/28
    surface finish, 8/26
  cubic boron nitride, 8/4
  diamond, 8/4, 8/30
  disc, 8/30
  lapping, 8/30–1
  Moh hardness scale, 8/4
  sand/quartz, 8/4
  SG, 8/17–18
  silicon carbide, 8/4, 8/30
  superabrasives, 8/6, 8/7–8
  three-body abrasion, 8/4
  two-body abrasion, 8/4
ABS (acrylonitrile- butadiene- styrene) polymers, 2/6, 2/24, 2/27
  laser-beam machining, 6/7
a.c. *see* Alternating current
Access protocol, networks, 13/18
Acetal resins, 2/24
Acrylics,
  acrylic rubbers, 2/15
  adhesives *see* Adhesives, toughened acrylics
  laser-beam machining, 6/7–8
Acrylonitrile-butadiene-styrene polymers *see* ABS
Actuators,
  hydraulic, 12/9–11, 12/12–13
  pneumatic, 12/9–12
Adaptive control, 11/18–20
  for quality, 17/16–17
ADC *see* Analogue-to-digital converters (ADC)
Adhesives,
  anaerobic,
    application of, 9/31
    characteristics, 9/29–30
    classification, 9/28
    cure mechanism, 9/29–30
    cylindrical component bonding, 9/30
    differential thermal expansion problems, 9/31
    gasketing, 9/31–2
    interference fits, 9/31
    preapplied, 9/30
    pre-heating, 9/31
    press fits, 9/31
    shrink fits, 9/31
    strength and torque, 9/30
    threadlocking, 9/30
    thread sealing, 9/32
  bond lines, 9/29
  cleavage, 9/28
  cyanoacrylates,
    activators, 9/32
    category, 9/28
    curing, 9/32
    humidity and moisture resistance, 9/32
    temperature resistance, 9/32
  dispensing equipment, 9/33
  environmental resistance, 9/29
  epoxies, 9/28, 9/33
  health and safety, 9/33
  high speed bonding, 9/32, 9/32–3
  hot melts, 9/28
  MOD approval, 9/29
  peel, 9/28
  phenolic, 9/28
  polyurethane, 9/28
  shelf life, 9/29
  silicones, 9/28
  solvent based, 9/28
  surface preparation, 9/29
  testing, 9/29
  toughened acrylics,
    application methods, 9/33
    applications, 9/33
    classification, 9/28
    cure mechanism, 9/32–3
    environmental resistance, 9/33
    temperature range, 9/33
  ultraviolet acrylics, 9/28, 9/32
AGVs (automated guided vehicles), 11/52, 11/65, 14/63
Alnico, 1/92
Alternating current (a.c.),
  see also Induction motors; Synchronous motors; Three-phase
  definition, 10/11
  frequency response of circuits, 10/16
  steady state, 10/13–15
Alternators, three-phase, 10/53
ALU (arithmetic and logic unit), 11/37
Alumina ceramics, 1/110–11, 1/116
Aluminium/aluminium alloys,
  aluminium oxide grinding abrasive, 8/4, 8/5
  BS codes, 1/63, 1/67–8
  BS specifications, 1/63, 1/66–8
  casting alloys, 1/63, 1/65–6, 1/67–8
  cold drawing, 4/97, 4/98
  continuous extrusion, 4/86
  corrosion resistance, 1/67
  Cosworth foundry, 1/66
  creep buckling, 1/68
  die casting, 1/66
  endurance limits, 1/68
  explosive forging, 4/125, 4/126
  exposure characteristics, 1/67
  fatigue crack growth, 1/68
  flow forming, 4/52
  forging, 4/63
  general properties, 1/59, 1/63
  hydrogen absorption, 1/66
  laser-beam machining, 6/7–8
  mashy state extrusion, 4/127, 4/128, 4/129
  powder fabrication, 1/66
  rake angles for machining, 5/15
  sand casting, 1/66
  static strengths, 1/68
  temperature characteristics, 1/69
  wrought alloys, 1/63–5, 1/66–7, 1/68
Aminoplastics, 2/25
Amplifiers,
  bandwidth, 10/23
  common mode rejection ratio (CMRR), 10/27, 11/33
  differential, 10/27
  distortion, 10/24, 11/34
  frequency response, 10/24
  gain, 10/24
  input impedance, 10/24, 11/34
  for instrumentation, 10/27
  negative feedback, 10/24
  noise, 10/24
  offset voltage and drift, 11/34
  operational,
    integrating, 10/27
    inverting, 10/25–6
    non-inverting, 10/26
    summing, 10/26–7
    unity gain, 10/26
  output impedance, 10/24
  positive feedback, 10/24–5
  stability, 10/24–5
  for transducers, 11/33–4
  types of, 10/23
Anaerobic adhesives *see* Adhesives, anaerobic
Analogue-to-digital converters (ADC),
  accuracy, 11/48
  conversion time, 11/48
  cost, 11/48

Analogue-to-digital converters
(ADC), *continued*
integrating (dual slope) type, 11/45, 11/47
parallel-conversion type, 11/47–8
resolution, 11/48
sample and hold (S/H), 11/45
staircase and comparator, 11/45, 11/46
successive-approximation type, 11/47, 11/48
Angular position transducer, 11/23–5
Anistropy, plastics, 2/43
Annealing, 1/30
see also Heat treatment
AOD *see* Argon oxygen degassing
AOQL (average outgoing quality level), 17/24–5
APPC (advanced program-to-program communications), 13/18
Application layer, OSI, 13/18
Application specific integrated circuits (ASICs), 10/43
APT (automatic programming tool), 14/20, 14/28, 14/56–7
AQL (acceptable quality level), 17/23
Aramid fibres, 1/96, 1/99
reinforcement, 2/6
ARCnet (attached resources computing), 13/18
Arc welding *see* Welding
Argon oxygen degassing (AOD), 1/13
Arithmetic and logic unit (ALU), 11/37
Arpanet network, 13/22
Artificial intelligence, 13/24
Asbestos,
fibres, 1/97
water-jet machining, 6/5
ASCII code, 11/36
ASICs (application specific integrated circuits), 10/43
Assel elongating,
metal rolling, 4/42, 4/45–6, 4/47–8, 4/49
strain and strain-rate relationships, 4/16–17
Assembly director, PCB assembly, 7/3
Astroloy, 1/86
Austenite *see* Steel
Austenitic stainless steel *see* Stainless steel
Auto-forge process, 6/69
Automated guided vehicles (AGVs), 11/52, 11/65, 14/63
Automated office, 19/5
Automatic programming tool (APT), 14/20, 14/28, 14/56–7
Availability, of equipment, definition, 18/5–6
Average outgoing quality level (AOQL), 17/24–5

Babbitt bearing metals, 1/92
Bainite steel, 1/24–5
Balestra type furnace, 4/79
Bandwidth, electrical, 10/16
Bar-code readers, 11/32–3, 14/86
for shop floor data collection, 14/10
Bar drawing, strain and strain-rate relationships, 4/16–17
Basic oxygen converter (BOC), 1/13
Batch buying, 15/52–3
Batch manufacture, 5/5–6
Bauds, 11/36
Bauschinger effect, 4/3
Bearings,
copper alloys for, 1/59, 1/61
lead/tin based, 1/92
Bellows, copper alloys for, 1/61
Beryllia ceramics, 1/111
Beryllium,
cold drawing, 4/97
nuclear application, 1/94
Bevel protractors, 17/35
Billet casters, 4/35
Bill of materials (BOM), 14/6, 14/8
Binary number system, 11/36
Blast furnace, 1/12
Blasting,
abrasives, 8/34
processes, 8/33–4
Bloom casters, 4/35
Blow moulding, plastics,
extrusion, 2/33
injection, 2/33
stretch, 2/33–4
BOC (Basic oxygen converter), 1/13
Bode stability criteria, 11/10–12
Bolts,
see also Threads
head forms, 9/4
Hi-bolt fastener, 9/5–6
load sensing, 9/5–6
RotaBolt, 9/5–6
BOM (bill of materials), 14/6, 14/8
Bonding *see* Adhesives
Boolean algebra *see* Logic systems
Boring machines, 5/3, 5/5
Boron carbide, 1/112
Boron filament fibres, 1/95–6, 1/97, 1/99
Boron nitrides, 1/112
Box pass rolling mill, 4/40
Brass,
cold drawing, 4/97
composition, 1/55, 1/56
corrosion resistance, 1/57–8
hot extrusion, 4/78
laser-beam machining, 6/7–8
machining properties, 1/57
strength, 1/57
stress corrosion, 1/58
Brazing,
applications, 9/26
methods, 9/26
Break-even analysis, 15/54–5
Bridge bearings, rubber, 2/38–40
Brinell hardness, 1/3–4
British Standards Institution (BSI),
assessment and certification, 17/10
BS 5750/ISO 9000, 17/12–13
BS 6001 sampling procedures and tables, 17/22–3
information services, 17/10
inspectorate, 17/12
Kitemark, 17/10, 17/12
preparation of standards, 17/10
registered firm of assessed capability mark, 17/12
safety mark, 17/12
technical help to exporters, 17/10
testing services, 17/12
Broaching machines, 5/4
Bronze,
aluminium bronze, 1/59
applications, 1/58
composition, 1/58
lead bronze, 1/58
phosphor bronze, 1/58
silicon bronze, 1/59
Brushless motors, 10/59, 12/5–6
BSI *see* British Standards Institution
Bulk deformation operation, metal shaping, 4/4
Bulk modulus, 1/3, 2/7–8
Burning tests, plastics, 2/17–18
Bus-bars, copper alloys for, 1/60
Bushes,
rubber, 2/38–40
for sleeves, copper alloys for, 1/61
Bus networks, 13/18
Bus structure, microprocessors,
address bus, 11/37
control bus, 11/37–8
data bus, 11/37
Butyl rubbers, 2/14

Cable,
balanced/unbalanced, 13/21
coax, 13/20
coaxial, 13/18
copper alloys for, 1/60
fibre optic, 13/20
twisted pair, 13/20
Cache, computer, 13/18
CAD *see* Computer aided design; Computer aided draughting
CADCAM *see* Computer aided design and manufacture
CADD *see* Computer aided design and draughting
CAE *see* Computer aided engineering
Calendering, plastics and rubbers, 2/35
Calmes elongator, metal rolling, 4/47
CAM *see* Computer aided manufacture
Cans, extrusion of, 4/75
Capacitors/capacitance,
in a.c. steady state, 10/13–14
air, 10/6
ceramic, 10/6
charging, 10/6
dielectric strength, 10/6–7
electrolytic, 10/6
mica dielectric, 10/6
paper dielectric, 10/6
parallel, 10/5
polycarbonate dielectric, 10/6
series, 10/5–6
transient switching of, 10/11–13
Capacity requirements planning (CRP), 14/8
CAPM (computer aided production management), 14/4
Carbon,
amorphous, 1/112
carbon-carbon composites, 1/102, 1/104
Carbon-fibre-reinforced polymer (CFRP), 1/95, 1/97, 1/98, 1/99, 1/100, 1/101, 2/6
fibre manufacture, 1/113
vitreous, 1/113
Carbon steel *see* Steel
Carrier sense multiple access (CSMA), 13/18
CASE (computer aided software engineering), 14/4
Casting,
see also Die casting; Investment casting; Moulds/moulding; Patterns; Sand casting
continuous steel casting, 1/13, 1/14
metal, 3/1–23
economics of, 3/3–4
plastics and rubbers,
centrifugal, 2/37
polymerisation, 2/36
steel, 1/50
titanium, 1/75
Cast iron,
austempered ductile, 1/53
austenitic, 1/54
blackheat malleable, 1/54
grey, 1/50–1, 1/53
high chromium, 1/54

high silicon, 1/54
martensitic white, 1/54
nodular graphite, 1/53
pearlite malleable, 1/54
spheroidal graphite (SG), 1/53
white, 1/50, 1/54
whiteheart malleable, 1/54
Cast-preform forging, 4/69
CAT (computer-automatic testing), 14/4
Cellulose plastics, 2/25
Cemented carbides, 1/107
Cementite steel, 1/21
Cemfil, 1/95
Ceramic cutting tools, 5/11
Ceramics,
see also Silicon carbide
alumina, 1/110–11, 1/116
aluminosilicates,
impermeable, 1/110
porous, 1/110
porous magnesium, 1/110
automotive components, 1/116
beryllia, 1/111
boron carbide, 1/112
boron nitrides, 1/112
compaction, 1/115
ductility, 1/114
elastic properties, 1/113
ferrites, 1/111
forsterites, porous, 1/110
glass ceramics, 1/113
hot pressing, 1/115
magnesia, 1/111
manufacturing, 1/115
mullite, 1/110
porcelains,
aluminous, 1/109
crystobalite, 1/109
pressed, 1/109
technical, 1/109
pyrolysis, 1/115
reaction bonding, 1/115
self bonding, 1/115
sialons, 1/112
silicon nitride, 1/112
sintering, 1/115
standards, 1/109
steatites,
low loss, 1/110
low voltage, 1/109
normal, 1/109–10
porous, 1/110
strength,
at elevated temperatures, 1/115, 1/117
fatigue, 1/115
long-term, 1/114
tensile, 1/114
testing methods, 1/114–15
titanates, 1/111
titania, 1/111
vitrification, 1/115
zirconia, 1/111, 1/116
Cermet cutting material, 5/10–11
CFRP (Carbon-fibre-reinforced polymer), 1/95, 1/97, 1/98, 1/99, 1/100, 1/101, 2/6
Chains, copper alloys for, 1/61
Charge, electrical, 10/5
Charpy V notch energy, 1/28, 1/45
Chatter, in machine tools, 5/11
Chlorinated polyethylene, 2/15
Chromium,
characteristics and applications, 1/93
explosive forming, 4/126
CIM (Computer integrated manufacture), 14/3–5, 15/50
Circuit breakers, 10/63
Circuits,
frequency response of, 10/15–16
integrated, 10/22
steady state a.c., 10/13–15
transient switching, 10/11–13
Classification and coding (CC), group technology, 15/22–4
Closed loop control, 11/4
CMOS (complementary metal oxide semiconductor) logic, 10/31
CMSR *see* Constant mean strain rate
CNC *see* Computer numerical control
Coaxial cable, 13/18
Cobalt/cobalt alloys,
Alnico, 1/92
applications, 1/92–3
dental applications, 1/93
in high-penetration radiography, 1/92
as a permanent magnet material, 1/92
Permender, 1/92
properties, 1/92
Cogging, forging preforming, 4/65
Coining, forging, 4/64
Cold drawing,
tube,
alloy steels, 1/21, 4/97
aluminium, 4/97
automation of plant, 4/104
beryllium, 4/97
brass, 4/97
CMSR die design, 4/101
coil drawing, 4/103–4
copper, 4/97
CRHS die design, 4/101
drawing load, 4/106
fixed plugs, 4/96–7, 4/103
floating plugs, 4/96–7, 4/103, 4/104
geometry factors, 4/19
homogeneous strain distribution, 4/102
lubrication, 4/97
mandel and plug strain and strain-rate relationships, 4/16
mandrel drawing stress, 4/106
molybdenum, 4/97
over long mandrels, 4/96–7
plug drawing stress, 4/106
post-nozzling annealing, 4/101
processes, 1/21, 4/96
product properties, 4/96
redundancy factors, 4/101
rotary swagers, 4/101
sinking process, 4/96, 4/102–3
sinking process stress, 4/105
stainless steel, 4/97, 4/103
tool design, 4/101–4
ultrasonic drawing, 4/104–5
ultrasonic vibratory systems, 4/97
wire and rod,
aluminium, 4/98
back-pull factor, 4/99
central burst, 4/99–100
chip formation, 4/99–100
copper, 4/98
dead-metal zones, 4/99–100
die design, 4/97, 4/99
dry process, 4/97–8
efficiency, 4/98–9
fine wires, 4/98
hydrodynamic lubrication, 4/98
industrial tools, 4/99
lubrication, 4/97, 4/98
mechanism, 4/98–9
processes, 1/21, 4/96
product properties, 4/96
strain and strain-rate relationships, 4/16–17
ultrasonic cleaning, 4/98
ultrasonic drawing, 4/98
ultrasonic vibratory systems, 4/97
wet process, 4/98
Commutators, copper alloys for, 1/60
Compacting, ceramics, 1/115
Compacting powders, 3/21–2
Comparators, 17/35
Complementary metal oxide semiconductor (CMOS) logic, 10/31
Complex operator j, 10/14
Composites,
see also Concrete, reinforced; Wood composites
boron-fibre-reinforced plastics, 1/99
carbon-carbon, 1/102, 1/104
cemented carbides, 1/107
continuous phase, 2/6
discontinuous fibre reinforced plastics, 1/102
dough-moulding compounds (DMC), 1/101
fibre reinforced glasses and ceramics, 1/104, 1/106
fibre-reinforced metals,
diffusion bonding, 1/104, 1/105
fabrication techniques, 1/104
materials, 1/104
properties, 1/105
squeeze casting, 1/104, 1/105
fibres, reinforcing,
Aramid, 1/96, 1/99
asbestos, 1/97
boron, 1/95–6, 1/97
carbon, 1/95, 1/97, 1/98, 1/99
cellulose, 1/97
Cemfil, 1/95
E glass, 1/95
glass, 1/99, 2/6
graphite, 1/95
Kevlar, 1/97
metal reinforcing, 1/96
nylon, 1/97
paraoriented aromatic polyamide, 1/96, 1/99
polypropylene, 1/97
short fibre properties, 1/103
silicon carbide, 1/96
steel, 1/97
general properties, 1/94–5
geometric classification, 1/95
interpenetrating, 2/6
laminar, 1/107
laminar structure, 2/6
matrix classification, 1/95
particulate, 1/107
polymer,
carbon-fibre-reinforced polymer (CFRP), 1/98, 1/99, 1/100, 1/101
chemical attack resistance, 1/98
continuous lamination, 1/96
elastic properties, 1/98, 1/99
epoxide resins, 1/96
fatigue, 1/98
glass-fibre-reinforced polymer (GFRP), 1/99
impact resistance, 1/98
laying, 1/96
moulding, 1/96
pultrusion, 1/96
strains to failure, 1/99
strength, 1/98, 1/99
thermophysical properties, 1/99

Composites,
polymer, *continued*
thermoplastic materials, 1/96
thermosetting resins, 1/96
winding, 1/96
reinforced thermoplastic materials (RTPs), 1/101
sheet-moulding compounds (SMC), 1/101
Compression moulding, plastics and rubbers, 2/28
Computer aided design (CAD), 11/37, 14/3, 14/12
see also Computer aided draughting
forging tools, 4/70
Computer aided design and draughting (CADD), 14/12
Computer aided design and manufacture (CADCAM), 14/12–27
centralised systems, 14/13, 14/14
displays for, 14/25–6
distributed systems, 14/13–15
electronic/electrical engineering, 14/16
fileservers for, 14/25
hard discs for, 14/25
input devices for, 14/25–6
mainframe computers for, 14/25
mechanical engineering, 14/16
minicomputers for, 14/25
mouse input, 14/26
multiple host systems, 14/13, 14/14
output devices for, 14/26–7
PCs for, 14/24–5
for printed circuit boards, 14/22–4
software,
applications, 14/16
communication, 14/16
operating system, 14/15–16
stand alone systems, 14/15
system hardware, 14/15
workstations, 14/25
Computer aided draughting (CAD),
see also Computer aided design
advantages/disadvantages, 14/16
future trends, 14/19
solids modelling, 14/17–19
surface modelling, 14/17
sytems operation, 14/16–17
wireframe modelling, 14/17
Computer aided engineering (CAE), 14/4, 14/12, 14/13, 14/19
Computer aided manufacture (CAM), 14/12
application and scope, 14/19–20
ergonomic considerations, 19/5
hardware, 14/22
software, 14/20–2
Computer aided production management (CAPM), 14/4
Computer aided software engineering (CASE), 14/4
Computer automatic testing (CAT), 14/4
Computer integrated manufacture (CIM), 14/3–5, 15/50
Computer numerical control (CNC),
see also Control systems; Machining centres; Turning centres
applications, 14/30
codes and formats, 14/50, 14/52, 14/53
controllers,
capabilities required, 14/46–7
closed loop feedback, 14/47–8
closed loop feedforward, 14/48
current developments, 14/48–9
hardware, 14/50
major functions, 14/49–50
open loop, 14/47
coolant control, 14/33
cutting spindles, 14/30
drive systems,
hydraulic drives, 14/31
pneumatic drives, 14/32
servodrive feedback, 14/30–2
stepping motors, 14/31
grinding, 8/12–13, 14/45–6
inspection in process, 14/33–4
interfacing, 13/21
machine input, 14/29
magnetic disc data, 14/51
magnetic tape data, 14/51
NC axis conventions, 14/30
post processing, 14/29
programming,
APT, 14/56–7
canned (fixed) cycles, 14/56
conversational, 14/57–8
graphical, 14/57
macros, 14/55
manual, 14/51, 14/54–6
methods outline, 14/52–4
subroutines, 14/55–6
sensing in process, 14/33–4
sheet metal punch presses, 14/44–5
slide movements,
rack and pinion, 14/32
recirculating ball and leadscrew, 14/32
slide position measurement,
inductosyns, 14/32
linear transducers, 14/32
positional accuracies, 14/32
rotary transducers, 14/32
swarf control, 14/33
tape,
codes and formats, 14/50, 14/52, 14/53
electromechanical readers, 14/50
optical readers, 14/50–1
pneumatic readers, 14/50
workheads, 14/30
Computers,
see also Computer numerical control (CNC); Databases; Microprocessors; Networks; Software
addressing schemes, 13/6–7
architecture,
basic structure, 13/3–4
classification, 13/4
artificial intelligence, 13/24
batch systems, 13/5
for CADCAM, 14/15, 14/24–5
cash memory, 13/6–7
CPU, 13/3–4
digiters, 13/15
displays, 13/10–11
expert systems, 13/24
eye position input, 13/15–16
in FMS systems, 14/67
gaze input, 13/15–16
graphic tablets, 13/15
interfacing,
analogue-to-digital converters (ADC), 11/45–8
CNC interface, 11/50
Darlington drivers, 11/41–2
digital-to-analogue converters (DACs), 11/42–4
flow valve control, 11/45
machine tools, 11/50
multiplexing, 11/48–50
opto-isolators, 11/42
pulse-width modulation (PWM), 11/44–5
robot control, 11/50–2
serial, 11/40–1, 13/21, 13/22
silicon controlled rectifiers (SCRs), 11/42
solid state relays, 11/42
thyristor power control, 11/45
joysticks, 13/13
keyboards,
DVORAK, 13/12
QWERTY, 13/11–12
knowledge based systems, 13/24
light pens, 13/14–15
maintenance applications, 18/15
manufacturing simulation, 14/11–12
memory managment schemes, 13/6–7
mice, 13/12–13
microcomputers, 13/5
minicomputers, 13/5
monitors, 13/10–11
multiuser systems, 13/5–6
operating systems, 13/5–6, 13/7
size classification, 13/4–5
speech recognition, 13/17
storage devices,
magnetic, 13/9–10
semiconductor, 13/8–9
terminals, 13/10–11
time sharing, 13/6
touch screens, 13/13–14
virtual memory, 13/7
voice input/recognition, 13/16–18
Computer simulation, metal flow, 4/72
Concrete, reinforced,
corrosion, 1/107
macro defect free (MDF), 1/106
pipes from, 1/106
polymer-impregnated, 1/106
Portland cement, 1/106
prestressed, 1/107
strength, 1/106
Weibull testing, 1/106
Concurrent engineering, 16/8–10
Condenser tubes, marine, copper alloys for, 1/62
Condition-monitoring,
applications,
absolute bearing vibration, 18/23
axial shaft vibration, 18/23
eccentricity, 18/23
oil debris, 18/24
pressure measurement, 18/24
relative shaft expansion, 18/23
relative shaft vibration, 18/23
rotational speed, 18/23
temperature measurement, 18/23–4
automatic monitoring, 18/22
description of, 18/21–2
manual periodic monitoring, 18/22
periodic monitoring, 18/22
Conductance, 10/3
Conform extrusion process, 4/86
Connectance, production system theory, 15/7
Constant mean strain rate (CMSR), 4/49
metal forming tools, 4/28, 4/29–30, 4/49, 4/90–1
Constant rate of homogeneous strain (CRHS), 4/90–1
metal forming tools, 4/28
Constitutive equation, plasticity, 4/17
Consumable vacuum arc process, 1/15
Consumer Protection Act, 17/6
Contacts, copper alloys for, 1/60
Continuous casting, steel, 1/13, 1/14, 1/15
Continuous line flow production, 15/19–20
Contracts, and quality, 17/6
Control systems,
see also Computer numerical control;

Direct numerical control; Frequency domain analysis; Stability; Time domain analysis
adaptive control, 11/18–20
cam-controlled machines, 12/3
cascade control, 11/17, 11/18
closed loop, 11/4, 12/3–4
computer based control, 11/18
continuous cycling stability method, 11/16
critical damping, 12/4
deadbands, 11/13
direct digital control, 11/17–18
disturbance sensistivity, 11/17
feedback, 11/4, 12/3
feedforward control, 11/17
frequency domain, 11/5
frequency domain analysis, 11/8–13
gain scheduling control method, 11/18–20
hardware, 12/5
hierarchical control, 11/21
integral control, 12/4
Laplace notation, 11/8–13
linear, 11/4
multilayer control, 11/21
multilevel control, 11/22
non-linear, 11/4
numerical control, 11/4
ON/OFF control, 11/13, 12/4
open loop, 11/4, 12/3
path control, 12/4
PID (proportional, integral and derivative) control,
derivative action, 11/15
integral action, 11/14–15
proportional action, 11/14
three-term controller, 11/16–17
Ziegler and Nichols rules, 11/15–17
process control, 11/4
proportional control, 11/13, 12/4
ramp input, 11/5
reaction curve stability method, 11/15–16
sampled-data control, 11/20
self-tuning control, 11/18–20
sequential control, 11/3
servomechanisms, 11/3
servomotor amplifiers, 12/7–8
sinusoidal input, 11/5
stability, 11/4–5, 11/9–12, 11/14–16, 12/4
step input, 11/5
time domain, 11/5–8
transport delay effects, 11/12–13
z-transforms, 11/21
Control valves,
hydraulic, 12/12–13
hydraulic servo, 12/12–13
pilot, 12/11
pneumatic, 12/11–12
poppet, 12/11
shuttle, 12/11
specifications, 12/10–11
symbols for, 12/11
Coordinate measurement machines, 17/42–3
Copolymers, 2/3
block, 2/5
random, 2/5
Copper/copper alloys,
alloy classes, 1/56
alloying additions, properties with, 1/55
alloys, 1/55–9
ASTM alloy standards, 1/57
for bearings, 1/59
chemical engineering applications, 1/61
cold drawing, 4/97, 4/98
condition coding, 1/57
conductivity, 1/55
continuous extrusion, 4/86
copper-lead alloy bearings, 1/59
copper-manganese alloys, 1/60
copper-nickel,
alloys, 1/59
explosive forging, 4/126
corrosion resistance applications, 1/62
electrical purposes guide, 1/60
general properties, 1/54–5
high conductivity alloys, 1/55, 1/56
for high temperature applications, 1/62
hot extrusion, 4/78
mechanical applications, 1/61
nickel silver alloys, 1/59
partial phase diagrams, 1/58
Correlation charts, 17/20
Corrosion protection coatings, zinc, 1/90–2
Costing,
cost accounting, 15/54
cost of operation, 18/7
cost of ownership, 18/7
internal rate of return (IRR), 18/11
life-cycle costing,
appraisal phase, 18/8
description, 18/7
life-cycle costs, 18/7
marginal analysis, 18/8–9
minimising costs, 18/10
practical difficulties, 18/10–11
process, 18/7–8
speculation phase, 18/8
trade-off models, 18/9–10
maintenance, 18/13–14, 18/15–18
payback, 18/10–11
quality related costs, 17/5–6
return on investment, 18/10
stock holding costs, 18/24–5
Cosworth foundry, aluminium, 1/66
Coulombs, 10/4
Counters *see* Logic systems
CPU (central processing unit), 11/37
Crack depth detection, electropotential, 17/56
Creep,
creep range, 1/4
creep rupture elongation, 1/4
Larson-Miller parameter, 1/4–5
plastics, 2/19–20, 2/43–4
stress to a certain creep strain, 1/4
stress to rupture, 1/4
CRHS *see* Constant rate of homogeneous strain
Croning shell moulding process, 3/20
CRP (capacity requirements planning), 14/8
Cryogenic applications, steel for, 1/45
CSMA (carrier sense multiple access), 13/8
Cubic boron nitride,
cutting tools, 5/11
grinding abrasive, 8/4
Cusum charts, 17/18–19, 17/20
Cutting,
gas/flame/oxygen, 9/21–2
laser, 9/23
plasma arc, 6/8
applications, 9/21
principle, 9/20–1
Cutting fluids, 5/13–14
Cutting tools *see* Tools/tool design
Cyanoacrylate adhesives, 9/28, 9/32
Cyclotol explosives for high-energy-rate metal forming, 4/115

DACs (digital-to-analogue converters), 11/42–3
Damp courses, copper alloys for, 1/62
Damping,
rubber components, 2/38
time domain analysis, 11/6
Darlington power drivers, 11/41–2
Databases,
database management systems (DBMS), 13/23
DBASE, 13/23
files, 13/23
records, 13/23
secretarial function, 15/55
Data link layer, OSI, 13/19
Datasheet PETN explosive, 4/115
Data transmission standards, 13/21
DBASE relational database, 13/23
DBMS (database management system), 13/23
dBs (decibels), 10/15
d.c. generators *see* Generators
d.c. motors *see* Motors
DDC *see* Direct digital control
Decibels (dBs), 10/15
Deep drawing sheet metal, 4/111–13
De Morgan's theorem, 10/30, 10/32, 10/35
Design, techniques to aid quality, 17/8
Design for manufacture and assembly (DFMA), 16/10
Design modifications, quality aspects, 17/7
Detonators for high energy metal forming, 4/113
DFMA (design for manufacture and assembly), 16/10
DH (Dortmund Horder) steel process, 1/13
Dial gauges, 17/34
Dial test indicators, 17/34
Diameter gauges, 17/37–8
Diamond, 1/112
cutting tools, 5/11
polycrystalline synthetic, 5/11
diamond passes rolling mill profile, 4/40
grinding abrasive, 8/4, 8/5, 8/7
Diaphragms, copper alloys for, 1/61
Die casting,
high pressure,
cold-chamber, 3/15
horizontal-chamber, 3/15–16
hot-chamber, 3/15
operating cycle, 3/16–17
low pressure, 3/14–15
Dielectric breakdown, 1/5
Dielectric loss, 1/5
Dielectric strength, 10/6–7
Dies, 4/8
see also Extrusion; Moulds/moulding
cold drawing, 4/97
geometry factor, 4/18
for high-energy-rate operations, 4/119
Diescher elongating, metal rolling, 4/42
Diffusion bonding, 4/127
Digital-to-analogue converters (DACs), 11/42–3
Digiters, computers, 13/15
Digitising tablets, CADCAM, 14/26
Diode, p-n junction, 10/17
Direct current motors *see* Motors
Direct digital control (DDC), 11/17–18
a.c. motor speed control, 11/58
d.c. motor speed control, 11/57–8
sampling frequency, 11/55–6
Stepper motor drive, 11/58
Direct numerical control (DNC), 14/28, 14/29
Discs, abrasive, 8/30
Disc storage,
floppy disk, 13/9
hard disk, 13/9–10, 14/25, 14/51

Displacement measurement *see* Transducers
Distortion, amplifiers, 10/24
Distributed numerical control (DNC), 14/28, 14/29–30
DMC (dough-moulding polyester resins), 2/25–6, 2/28
DNC *see* Direct numerical control; Distributed numerical control
Dortmund Horder (DH) steel process, 1/13
Downtime, equipment, 18/6
Drawing *see* Cold drawing
Drilling machines, 5/5
Duplex stainless steel, 1/12, 1/38

Earth leakage protection, 10/63
EBQ (economic batch quantity), ordering, 15/44–6
ECAD (electronic CAD), 14/4
ECH (electrochemical honing), 6/5
ECM *see* Electrochemical machining; Electrolytic grinding
Economic batch quantity (EBQ), ordering, 15/44–6
Economics,
  machining, 5/20–5
  welding, 9/26–7
ECT (electrochemical turning), 6/5
Eddy current loss, 10/10
EDM *see* Electrical discharge machining
Efficiency, metal forming, 4/18
E glass, 1/95
Elasticity *see* Plasticity; Stress-strain relationship
Electrical discharge machining (EDM), 6/3–4
  accuracy, 6/3
  applications, 6/4
  metal removal rates, 6/3
  surface finish, 6/3
  tool materials, 6/3
  tool ware, 6/3
Electrical energy, 10/4
Electrical networks, 10/4
Electrical power, 10/4
Electrical safety *see* Safety, electrical
Electric arc steel process, 3/4
Electric steels, 1/28–9
Electrochemical honing (ECH), 6/5
Electrochemical machining (ECM), 6/4
  surface finish, 6/4
Electrochemical turning (ECT), 6/5
Electro-erosion *see* Electrical discharge machining
Electrogalvanising, 1/92
Electrolytic belt grinding, 6/5
Electrolytic end milling, 6/5
Electrolytic grinding, 6/4–5
Electrolytic polishing, 6/5
Electromagnetic systems, 10/6–7
Electromotive force, 10/4
Electron-beam machining,
  advantages and limitations, 6/7
  applications, 6/7
  principle, 6/7
Electron beam steel process, 1/15
Electronic CAD (ECAD), 14/4
Electronic logic design *see* Logic systems
Electronic manufacture *see* Printed circuit boards
Electropotential crack depth detection, 17/56
Electroslag remelting (ESR), 1/15, 1/17, 1/21
Electrostatic systems, 10/4–5
Electro-stream, 6/5
Embossing, forging, 4/64
e.m.f (electromotive force), 10/7–8
  self induced, 10/8–9
Encoders, optical,
  absolute, 12/19–20
  Gray code, 12/20
  interferential scanning, 12/18
  Moire fringes, 12/17
  principle of operation, 12/17
  signal processing, 12/19
Energy, electrical, 10/4
Epichlorhydrin rubbers, 2/15
Epoxide (epoxy) resins, 2/26
Epoxy adhesives, 9/28, 9/33
EPROM (erasable programmable read only memory), 11/38, 11/53, 11/54, 13/9
Equilibrium technique, metal forming modelling, 4/19–20
Equipment,
  availability, 18/5–6
  downtime, 18/6
  maximum time to repair, 18/6
  mean time between failures (MTBF), 18/6
  mean time to repair, 18/6
  operating cycle, 18/6
  reliability, 18/6
  usage rate, 18/6
Equiverse type furnace, 4/79
Erasable programmable read only memory (EPROM), 11/38, 11/53, 11/54, 13/9
Ergonomics,
  automated office, 19/5–6
  computer aided manufacture, 19/5
  economics of, 19/6–7
  manual assembly, 19/6
  relevance to engineering, 19/3–4
  robotics, 19/4–5
ESR (electroslag remelting), 1/15, 1/17, 1/21
Ethernet network, 11/64, 13/19, 13/23
Ethylene, 2/3
Ethylene-acrylic rubbers, 2/15
Ethylene-propylene copolymers, 2/14
Ethylene-propylene-diene terpolymers, 2/14
Eurocard, 11/58–9
Euronet network, 13/22–3
Eutectoid steel, 1/23
EVA (ethylene-vinyl acetate), 2/22
Exotic rubbers, 2/15
Expert systems, 13/24
Explosive forging, 4/124–6
Explosive hardening, 4/126
Explosives, metal forming,
  cord, 4/116
  cyclotol, 4/115
  Datasheet, 4/115
  low (deflagratiing), 4/113, 4/115–17
  Metabel, 4/115
  primary high, 4/113, 4/115–17
  secondary high, 4/113, 4/115–17
  TNT, 4/116
Explosive welding, 4/121
  rod arrays, 4/82
  tubes, 4/105
Extrusion,
  aluminium, 4/78
  backward, 4/74–5
  basic operations, 1/21, 4/73–8
  brass, 4/78
  cans, 4/75
  circular to elliptical profile, 4/94
  cold, 4/80–6
  composite metal components, 4/88
  Conform process, 4/86, 4/89
  constant mean strain rate, 4/90–1
  constant rate of homogeneous strain, 4/90–1
  continuous, 4/86
  copper, 4/78
  dead-metal zone, 4/77, 4/80
  defects, 4/75–7
    cavitation, 4/85, 4/86
    core fracture, 4/84
    die contact fracture, 4/77
    fish skin, 4/84, 4/86
    inhomogeneous deformation, 4/77
    internal fracture (central bursting), 4/77
    sleeve fracture, 4/84
    surface cracking, 4/77
  die design, 4/88–96
    CAD, 4/94
    die materials, 4/96
    flat-faced dies, 4/92
    profiled dies, 4/93
    smoothed face dies, 4/92
  flow of metal, 4/88–96
  forward, 4/73–5
  general concepts, 4/73
  geometry factors, 4/19
  glass as lubricant, 4/73, 4/79
  high energy, 4/84
  hydrostatic, 4/80–1, 4/83
    basic, 4/81
    billet augmented, 4/81
    differential, 4/81
    lubrication, 4/81
    product augmented, 4/81
  impact, 4/84, 4/86
  implosive welding as pre-operation, 4/82
  loose array problems, 4/82
  losses, shear and frictional, 4/86
  magnesium, 1/78
  mashy state, 4/127–8
  metal, 4/10
  nickel alloys, 4/78
  plastic flow of metal, 4/10
  presses for, 4/73
  ram profiles, 4/85–6
  redundant deformation, 4/90
  rubbers and plastics,
    dies, 2/32
    multi-screw, 2/32
    single screw, 2/31–2
  section, 4/88
  shearing distortion, 4/90
  side, 4/74–5
  square-die processing, 4/73
  steels, 1/21, 4/78
  strain and strain-rate relationships, 4/16
  titanium alloys, 4/78
  tubes,
    billet heating, 4/79
    billet preparation, 4/78
    lubricant removal, 4/80
    lubrication, 4/73, 4/79
    pilot holes, 4/79–80
    seamless, 4/78
    stainless steel, 4/78, 4/80
    staving, 4/80
    Ugine-Sejournet process, 4/79, 4/80
  work redundancy factor, 4/85
Eye position input, computers, 13/15–16

Factory flow analysis (FFA), 15/23, 15/26–7
  see also Group technology
Failure mode, effect and criticality analysis, 17/9–10, 17/11
Faradays law, magnetic systems, 10/7–8
Farads, 10/5
Fasteners,
  see also Bolts; Nuts; Screws; Threads
  automatic insertion, 9/3–4
  Heatserter, 9/3
  part punching, 9/4

Pemserter, 9/3
plastic, 9/8, 9/9
PSM, 9/7
requirements of, 9/3
rivets, 9/8, 9/9–10
self sealing, 9/8, 9/9
spring steel, 9/8
supplier list, 9/10
threaded, 9/9–5
threadlocking, 9/6–7
Fatigue,
fatigue endurance, 1/3
fatigue endurance limit, 1/4
fatigue life for p% survival, 1/4
fatigue limit, 1/4
fatigue notch factor, 1/4
fatigue notch sensitivity, 1/4
FEA *see* Finite element analysis
Feedback control, production management, 15/6–7
Feeler gauges, 17/34
FEM (finite element mesh), 14/19
Ferrite,
see also Steel
as a ceramic, 1/111
Ferritic stainless steel *see* Stainless steel
FET (field effect transistor), 10/21–2
FFA (factory flow analysis), 15/23, 15/26–7
Fibreglass,
see also Composites
water-jet machining, 6/5
Fibre optics, 13/19
Fibre reinforcement *see* Composites
Field effect transistor (FET), 10/21–2
File servers,
CADCAM, 14/15, 14/25
CAM, 14/22
LAN, 13/19
Filters,
digital, 11/35–6
high pass, 11/35
low pass, 11/34–5
for transducers, 11/34–6
Finance,
break-even analysis, 15/54–5
cost accounting, 15/54
as management function, 15/54
Fine blanking metal, 4/108
Finite element analysis (FEA), 14/19
Finite element mesh (FEM), 14/19
Finite element process, modelling metal formation, 4/23
Fishbone diagrams, 17/18, 17/19
Flame retardancy, plastics, 2/17–18
Flanging, tubes or blanks, 4/110
Flexible manufacturing,
advantages and disadvantages, 14/60–1
applications, 14/61
cells (FMC), 14/58–9
control systems,
functions, 14/66–7
hardware, 14/67
definition, 14/58–60
fixtures, 14/63–4
flexible transfer lines, 14/59
machining centres, 14/62
pallet use, 14/63
planning,
design factors, 14/72
project management, 14/72–3
strategy, 14/68, 14/72
simulation, 14/68
systems (FMS), 11/64–5, 14/28, 14/58, 15/20
tool changing, 14/65
tool management,
cell centred, 14/64
machine centred, 14/64
system centred, 14/65
transportation methods, 14/65
trends, 14/61
turning machines, 14/63
Flexural strength, 1/3
Floturning, 4/51, 4/53
Flow forming tubes, 4/51–3
Flowmeters *see* Transducers
Fluid power actuators *see* Actuators
Fluoroplastics, 2/25
Fluororubbers, 2/15
Flux, 10/3
FM *see* Flexible manufacturing
FMC *see* Flexible manufacturing, cells
FMS *see* Flexible manufacturing, systems
Foam,
mouldings, 2/44
polyurethane, 2/37
Forging,
aluminium, 4/63
auto-forge process, 6/69
barrelling, 4/57
basic concepts, 4/56–7
blocker dies, 4/60
blocking, 4/60, 4/62
bulging, 4/60
carbon tool steel, 1/19
cast-preform, 4/69
closed-die, 4/59–60, 4/62–5
cogging preforming, 1/20, 4/65
coining, 4/64
cold working, 4/56, 4/62–4
die steel, 1/19
drawing preforming, 4/65
embossing, 4/64
flow deformation, 4/57
force equilibrium calculations, 4/72
fullering preforming, 4/65
heading preforming, 4/65
hollow forgings, 1/21
hot forging, 4/56
hubbing preforming, 4/65
impression-die, 4/59–61
flash, 4/60, 4/62
operational speeds, 4/60, 4/62
typical sequence, 4/61
isothermal, 4/70
lead, 4/63
load estimation, 4/72–3
magnesium, 1/78
open-die, 1/20, 4/57–9
forces and pressures, 4/57–9
orbital, 4/69
piercing, 4/60, 4/62, 4/64
pilger mill, 4/66–9
powder forming, 4/64–5
rotaform process, 4/69
rotary, 4/66
rotary tube, 4/66–9
tool design, 4/67
simulation of metal flow, 4/72
sinter forging, 4/64–5
steel, 1/19, 1/20–1
steel tungsten, 1/19
strain hardening, 4/56
stress calculation, 4/72
temperatures for, 4/56
tin, 4/63
titanium, 1/75
tooling, 4/70–2
CAD approach, 4/70
lateral shape factor, 4/70
longitudinal shape factor, 4/70
shape difficulty factor, 4/70
tubing,
seamless, 4/66–7
welded, 4/68
tungsten, 1/19
upsetting, 4/57–9
warm working, 4/63
zero draft tooling, 4/62
Forming metal,
see also Casting; Cold drawing; Extrusion; Forging; High-energy-rate metal forming; Moulds/moulding; Plastic flow of metal; Sheet metal forming; Tools/tool design
cold processes, 4/13
distortion, 4/12
efficiency, 4/20
flow characteristics, 4/12
formability comparison, 4/116
hot processes, 4/13
mashy state operations, 4/127–9
mathematical modelling, 4/19–23
operational temperatures, 4/4–5
plastic flow origin and nature, 4/3–11
plasticity theory, 4/13–27
process assesment, 4/11–13
process classification, 4/4
redundancy, 4/18, 4/20
rolling processes, 4/35–56
simulation of metal flow, 4/23–7
Strain, useful/redundant, 4/12
superplastic, 4/126–7
and mashy state, 4/126–30
Foundries *see* Casting
Fracture toughness,
elastic plastic, 1/4
plain strain, 1/4
stress intensity for stress corrosion, 1/4
Frame, token ring packet, 13/19
Frequency domain analysis,
first order systems, 11/9
second order systems, 11/9
Frequency response,
amplifiers, 10/24
circuits, 10/15–16
Friction, metal forming, 4/12
Frictioning, plastics and rubbers, 2/35
Fullering, forging preforming, 4/65
Furnaces,
billet heating,
Balestra type, 4/79
Equiverse type, 4/79
gas fired, 4/79
induction, 4/79
salt type, 4/79
Selas radiant heat slot, 4/79
blast, 1/12
Fuses, 10/63

Gain scheduling control method, 11/18–20
Gamma radiation, 17/47–8
Gasketing, with adhesives, 9/31–2
Gas turbines materials *see* Nickel/nickel alloys
Gates, logic *see* Logic systems
Gates, sand casting, 3/9–12
Gateway, network, 13/19
Gauge blocks, 17/35
Gauges *see* Measurement/measuring instruments
Gaze input, computers, 13/15–16
Gears,
backlash, 12/8–9
copper alloys for, 1/61

Gears, *continued*
double helical, 12/8
helical, 12/8
herringbone, 12/8
hobbing machines, 5/5
rack and pinion, 12/8
screw-nut systems, 12/8–9
shapers/planers, 5/3–4
spur, 12/8
stick-slip effect, 12/8–9
worm, 12/8
Generators, d.c.,
armature e.m.f., 10/44–5
armature torque, 10/45
back e.m.f., 10/45
carbon brushes, 10/43
commutators, 10/43
compound wound, 10/46, 10/47
efficiency, 10/51
lap winding, 10/45
seperately excited, 10/45–6
series-wound, 10/46, 10/47
shunt-wound, 10/46, 10/47
two pole, 10/44
wave winding, 10/45
Generator windings, copper alloys for, 1/60
Germaium, in integrated circuits, 1/94
GFRP (glass-fibre-reinforced polymer), 1/99, 2/6
Glass,
alkali-lead, 1/113
aluminosilicate, 1/113
borosilicate, 1/113
fused quartz, 1/113
fused silica, 1/113
glass ceramics, 1/113
laminated, 1/107
as metal forming lubricant, 4/32
sodalime silicate, 1/113
vitreous silica, 1/113
water-jet machining, 6/5
Glass-fibre-reinforced polymer, 1/99, 2/6
Gold, applications, 1/94
Grade, quality, 17/3
Graphics display monitors, CADCAM, 14/25–6
Graphic tablets, computers, 13/15
Graphite, 1/112, 1/113
Gray code, 11/36–7
Grinding/grinding wheels,
see also abrasives
balancing,
automatic, 8/7
coarse and fine, 8/7–8
diamond, 8/7–8
dynamic, 8/7
static, 8/7, 8/8
superabrasives, 8/8
belt grinding, 8/29–30
bonds,
metal, 8/6
resinoid, 8/6
vitrified, 8/6
centreless grinding, 8/15–17
computer numerical control (CNC), 8/12–14
cooling/coolants, 8/11, 8/12
cracks, 8/11
creep feed grinding, 8/12
cutting off, 8/17
cylindrical grinding,
external, 8/13, 8/16
internal, 8/13–15
dressing, crush, 8/7
energy input, 8/10–11
forces, 8/11–12
grades, 8/6
grit sizes, 8/5–6
heat effects, 8/11
mounting, 8/6–7
plunge grinding, 8/12
ratio, 8/11
self sharpening, 8/10
SG abrasives, 8/17–18
structure, 8/6
surface grounding, 8/12–13
surface measurement,
stylus, 8/9–10
taper sectioning, 8/9–10
theory, 8/9–10
tool and cutter grinding, 8/17
wear,
attritious, 8/11
bond post-rupture, 8/11
fracture, 8/11
work clamping, 8/13
Grinding machines, CNC,
design, 14/45
machine control and programming, 14/46–7
Group analysis *see* Group technology
Group technology (GT) production, 15/20–2
advantages, 15/22–4
assembly groups, function, 15/32
classification and coding (CC), 15/22–4
component processing groups, 15/29
factory flow analysis (FFA), 15/23, 15/26, 15/27
group analysis,
division into modules, 15/27–8
implementation and problems, 15/27
modules, 15/31, 15/32
special plant lists (SPL), 15/28
group assembly examples, 15/35
groups and families, 15/29
planning assembly groups, 15/32
production flow analysis (PFA), 15/23, 15/26, 15/27
service groups, 15/21
Gunmetal, 1/58

Hall effect sensors, 11/29
Hardening *see* Heat treatment
Hardness,
Brinell hardness, 1/3–4
Rockwell hardness, 1/3–4
rubber, 2/7–9
Vickers hardness, 1/3–4
Hastelloys, 1/81
HDPE (high density polyethylene), 2/22, 2/27
Heading, forging preforming, 4/65
Health-and-safety,
adhesives, 9/33
Health and Safety at Work Act, 17/6
Heat properties, plastics, 2/16–17
Heat treatment, steel,
allotropy of iron, 1/21
annealing, 1/30
austempering, 1/29
carbonitriding, 1/35
carburising, 1/31, 1/35
case hardening, 1/31–5
continuous cooling, 1/25
eutectoid line, 1/23
ferritic steel constitution, 1/21
flame hardening, 1/31
gas nitriding, 1/35
hardenability of steel, 1/30
iron-carbon composition diagram, 1/22–3
iron-carbon phase diagram, 1/18, 1/22
isothermal annealing, 1/30
isothermal transformations, 1/24
Jominy end-quench test, 1/30
liquid nitriding, 1/35
martempering, 1/30
nitriding, 1/35
quenching, 1/29
solid solutions, 1/21
surface hardening, 1/31
tempering, 1/29
time-temperature transformation (T-T-T), 1/24
Helical gear cutting, 5/4
Hencky's equation, modelling, 4/22
HERF *see* High-energy-rate metal forming
Hexadecimal number system, 11/36
High-energy-rate metal forming,
applications, 4/111, 4/113, 4/118–19
die design, 4/119
direct transmission, 4/117
electromagnetic forming, 4/118, 4/120–1
energy sources, 4/113–18
explosive forging, 4/124–6
explosive hardening, 4/126
explosive welding, 4/121
fabrication (HERF), 4/6, 4/10–11
high explosives for, 4/113, 4/115–17
hydroelectric forming, 4/117, 4/119–20
exploding wire, 4/117
underwater discharge, 4/117
low (deflagrating) explosives, 4/113, 4/115–17
plastic flow of metal, 4/10–11
powder compaction, 4/121–2
processes, 4/117
transmission through a medium, 4/117
High polymers, 2/3
Hill and Tupper method for plasticity strain equations, 4/18
Hinges, copper alloys for, 1/62
HIPS (high-impact polystyrene), 2/24
Hole-depth gauges, 17/37
Honing,
see also Superfinishing
abrasives, 8/19
basic principles, 8/18–19
CBN hones, 8/19
diamond hones, 8/19
fixtures for, 8/20, 8/21
Honamould tools, 8/19–20
machine configurations, 8/20
plateau honing, 8/23
sizing systems, 8/20, 8/22
surface characteristics, 8/20–3
Hot dip galvanising, 1/90, 1/92
Hot pressing, ceramics, 1/115
Hubbing, forging preforming, 4/65
Huber-Mises criterion of plastic flow yield, 4/18
Hydraulic systems,
CNC machines, 14/31
PLC control, 11/62–3
Hydraulic valves and actuators, 12/9–11, 12/12–13
Hydroelectric high-energy-rate metal forming, 4/117–18
Hydro/hydromechanical metal forming, 4/111
Hydromech, 4/111
Hydrostatic extrusion, 4/80–1, 4/83
Hysteresis, magnetic circuits, 10/9–10

IDEF management recording system, 15/12
Identification systems, automatic, 14/86–7
IEEE–488 (GPIB) Hewlett-Packard interface bus, 11/41, 11/58–9, 13/21

Iliums, 1/81, 1/83
Impact testing, 1/4
Impedance, 10/14
Implosive welded rod arrays, 4/82
Inconels, 1/81–2, 1/83
  explosive forging, 4/126
Inductance,
  mutual, 10/9
  transient switching, 10/11–13
Induction furnaces, 4/79
Induction motors,
  braking,
    dynamic, 10/55
    plugging, 10/55
  capacitor single-phase, 10/57
  shaded pole single-phase, 10/57
  slip ring, 10/54
  speed control,
    pole number changing, 10/56
    rotor resistance changing, 10/56
    solid state variable-frequency, 10/56
    stator voltage reduction, 10/56
  squirrel cage 3 phase, 10/54
  starting,
    autotransformer, 10/55
    rotor resistance, 10/55
    star-delta, 10/55
  torque characteristics, 10/54–5
  universal a.c./d.c., 10/57
Inductosyns, 11/25, 12/17, 14/32
Ingot casting, 1/13, 1/15
Injection moulding, plastics and rubbers, 2/27, 2/29–31
Input/output, microprocessors,
  dedicated port addressed, 11/39
  direct memory access (DMA), 11/39
  memory mapped, 11/39
  support chips, 11/39
Instrumentation *see* Transducers
Insulation, electrical, 10/62–3
Integrated circuits, 10/22
Integrated design, quality aspects, 17/7–10
Interfaces/interfacing *see* Computers, interfacing
Internal rate of return (IRR), 18/11
Inventory,
  holdings for maintenance, 18/24
  reorder quantities, 18/25–6
  safety stock, 18/25
  stock holding costs, 18/24–5
Investment casting,
  features, 3/17
  moulds, 3/17, 3/18–19
  patterns, 3/17, 3/18
  Shaw process, 3/17–19
  slurry, 3/17
Iridium, 1/94
Iron *see* Cast iron
Iron carbide, 1/21
IRR (internal rate of return), 18/11
ISO (International Standards Organisation), 13/19
Isolation, electrical, 10/63
ISO metric thread system, 9/4
Isostatic pressing, 3/21–3
Isothermal forging, 4/70
Isothermal transformation, steel, 1/24

JFET (junction field effect transistor), 10/22
JIT *see* Just-in-time
Jominy end-quench test, steel, 1/30
Joystick, computer, 13/13
Junction field effect transistor (JFET), 10/22
Just-in-time (JIT),
  batch sizes, 16/8
  bills of materials, 16/8
  lead times, 16/8
  organisation, 16/8
  philosophy, 16/8
  production, 15/15
  production control, 15/46, 15/50
  product structures, 16/8

Karnaugh map, 10/36–7, 10/38
Kevlar, laser-beam machining, 6/7
Keyboards,
  for CADCAM, 14/26
  DVORAK, 13/12
  QWERTY, 13/11–12
Kirchoff's laws, 10/4, 10/10–11
Kitemark, BSI, 17/10, 17/12
Knowledge based systems, 13/24

Laminates, 2/6
  fibre-reinforced, 2/36–7
Languages, computer *see* Software
LANs *see* Local area networks
Laplace notation,
  first order systems, 11/9
  Laplace transform, 11/8, 11/9
  second order systems, 11/9
Lapping,
  Aluminium oxide abrasive, 8/30
  diamond abrasive, 8/30
  machine layouts, 8/32
  non-embedding compounds, 8/30–1
  pearlite cast iron discs, 8/31
  processes, 8/31, 8/33
  silicon carbide abrasive, 8/30
Larson-Miller parameter, 1/4–5
Laser-beam,
  machining, 6/7–8, 9/23
    applications, 6/7–8
  welding, 9/22–3
Lathes, 5/3
  see also Turning centres
Law, quality and the, 17/6–7
Lead/lead alloys,
  as bearing materials, 1/92
  corrosion characteristics, 1/92
  creep strength, 1/92
  forging, 4/63
  nomenclature, 1/92
  solder, 1/92
  standards, 1/92
  toxicity, 1/92
Leather,
  laser-beam machining, 6/7
  water-jet machining, 6/5
Length bars, 17/35
Levy-Mises rule of flow, plasticity, 4/19
Liability, vicarious, 17/6
Life-cycle costing *see* Costing
Lightning conductors, copper alloys for, 1/62
Light pens,
  CADCAM, 14/26
  computers, 13/14–15
LLDPE (linear low-density polyethylene), 2/22
Local area networks (LANs), 13/22
  bridge, 13/8
  with PLCs, 11/64
Locking, network application files, 13/19
Logic systems,
  astable circuits, 10/42
  Boolean algebra,
    applications, 10/33–5, 10/36–7
    De Morgans theorem, 10/30, 10/32, 10/35
    exclusive OR, 10/30
    expression minimisation, 10/37–8
    logical AND/NAND/NOR/NOT/OR, 10/29–30, 10/32–5
  counters,
    asynchronous, 10/41
    ripple, 10/41
    synchronous, 10/41
  flip flops,
    D type, 10/41
    JK type, 10/39–40, 10/41
    SR type, 10/39
  gate symbols, 10/31
  Karnaugh map, 10/36–7, 10/38
  latches, 10/35–6
  monostables, 10/42
  registers, 10/40–1
  sequential logic circuits, 10/39–40
  timers, 10/42
  tri-state logic, 10/38–9
  truth tables, 10/30, 10/32–3
Lost-wax casting process *see* Investment casting
Lubrication, metal forming tool design, 4/31–2

Machine part production, rolling, 4/53–4
Machine-tools, computer interfaces, 11/50
Machining,
  see also Grinding/grinding wheels; Honing; Machining centres; Tools/tool design; Turning centres; Ultrasonic machining
  boring, 5/5
  broaching machines, 5/4
  chatter, 5/11, 5/25
  chip breakers, 5/3
  chip formation, 5/11–14
  costs, 5/21
  cutting fluids, 5/13–14
  drilling, 5/5
  economics of, 5/20–5
  electrical discharge, 6/3–4
  electrochemical machining, 6/4
  electro-erosion, 6/3–4
  electron beam, 6/7
  forces for, 5/14–17
  gear hobbing machines, 5/5
  gear shapers/planers, 5/3–4
  heat generated, 5/13
  hot machining, 5/17
  idle time, 5/21, 5/23
  Laser-beam, 6/7–8
  lathes, 5/3
  lubrication, 5/14
  machining costs, 5/21–5
  milling forces and power, 5/15–17
  milling machines, 5/4, 5/7, 5/9
  planing machines, 5/3
  plasma arc, 6/8
  production costs, 5/21–5
  rake angles, 5/7–10, 5/14–15
  set-up time, 5/21
  shaping machines, 5/3
  sintered cutting inserts, 5/3
  slotting machines, 5/3
  spark machining, 6/3–4
  surface finish, 5/17
  swarf disposal, 5/5, 5/14
  turning forces and power, 5/14–15
  vertical boring machines, 5/3
  water-jet, 6/5
Machining centres,
  in flexible manufacturing, 14/62
  horizontal, 14/35
  tooling,

Machining centres,
tooling , *continued*
automatic tool changers, 14/38
linear chain carousel, 14/39
rotary carousel, 14/38
types available, 14/34–5
vertical, 14/35, 14/36
workholding,
fixtures, 14/36–7
indexing tables, 14/37
pallet systems, 14/35–7
MAC (Media access control), 13/19
Magnesia ceramics *see* Ceramics
Magnesium/magnesium alloys,
applications, 1/79
casting, 1/78
physical properties of, 1/79
cold working, 1/78
compressive properties, 1/76–7, 1/80
corrosion, 1/78, 1/79
designation, 1/78
extrusion, 1/78
fatigue endurance, 1/76–7, 1/80
forging, 1/78
hardness, 1/76–7
high temperature strength, 1/78–9
standards, 1/76–7, 1/78, 1/80
tensile properties, 1/76–7, 1/80
wrought, 1/80
Magnetic cores, steel for, 1/28–9
Magnetic particle testing,
detectors, 17/54
magnetisation techniques,
coil shot, 17/54, 17/55
current flow, 17/54
flux flow, 17/54, 17/55
principle, 17/52, 17/54
viewing, 17/54
Magnetic storage *see* Discs; Tape
Magnetic stripes, 14/86
Magnetic theory,
circuits, 10/9–11
fields, 10/6–7
Maintenance,
see also Condition-monitoring; Inventory
computer use for, 18/15
contracts, 18/18
control, 18/14–15
costs,
direct, 18/16
engineering staff, 18/13
indirect, 18/16
parts, 18/13–14
equipment availability, 18/5–6
equipment downtime, 18/6
equipment operating cycle, 18/6
equipment reliability, 18/6
equipment usage rate, 18/6
indices,
availability, 18/14
effectiveness, 18/14
prime maintenance index, 18/14
managers role, 18/4
maximum time to repair, 18/6
modifications, 18/18
planning, 18/14–15
preventative,
failure patterns, 18/20
maintenance type classification, 18/19–21
routines, establishing, 18/20–1
pricing, 18/17–18
profit centre concept, 18/15–16
refurbishing, profit from, 18/18–19
revenue, 18/16–17
and terotechnology, 18/4
Management,
see also Finance; Manufacturing; Marketing; Personnel; Production; Production control; Purchasing; Secretarial
of assets, 18/3
computer-integrated manufacturing, 15/11–12
control, 15/3
direction, 15/3
function division, 15/3, 15/7–8
GRAI pyramid method, 15/9–11
hierarchical levels, 15/8–11
IDEF recording system, 15/12
integrated manufacturing, 15/11–12
MODUS recording system, 15/12
organisation, 15/13
personnel control, 15/13–14
planning, 15/3
policy, 15/12–13
product design function, 15/14–15
production planning, 15/15–18
quality, 17/4
strategic, 15/12–13
system recording, 15/12
systems theory,
black box, 15/7
connectance, 15/7, 15/8, 15/9, 15/10
feedback control, 15/6–7
input variables, 15/6
monitoring, 15/7
output variables, 15/6
parameters, 15/6
system design, 15/7
system design parameters, 15/6
Mandrell mill system, metal rolling, 4/43
Manfred-Weiss process, metal rolling, 4/45
Mannesmann cycle, metal rolling, 4/42, 4/47
Manufacturing,
see also Production
computer-aided manufacturing (CAM), 14/3
computer based simulation, 14/11–12
computer-integrated (CIM), 14/3–5, 15/11–12
concurrent engineering, 16/8–10
control systems, 14/5–12
design for manufacture and assembly (DFMA), 16/10
environment, 1980s, 16/4–5
environment changes,
inventory policy, 16/3
manpower levels, 16/3
organisation structure, 16/3
product designs, 16/3
production methods, 16/3
production technology, 16/3
product ranges, 16/3
quality assurance, 16/3
quality control, 16/3
skills, 16/3
volume levels, 16/3
group technology, 16/4–5
integrated, 15/11–12
manufacturing information systems (MIS), 14/4
organisation structure,
future, 16/6–7
past, 16/5–7
resources planning, 14/7–10
strategies,
1980s, 16/4–5
future, 16/5
key objectives, 16/5
world class manufacturing (WCM), 16/7
Manufacturing automation protocol (MAP), 13/19
Manufacturing-resource planning (MRP II), 14/7–8
MAP (manufacturing automation protocol), 13/19
Maraged steels, 1/12, 1/45–6, 1/47
Marble, water-jet machining, 6/5
Marginal analysis of costs, 18/8–9
Marketing,
business forcasting, 15/36
environment changes,
competition, 16/3, 16/4
markets, 16/3
price, 16/3
product ranges and life, 16/3
quality expectation, 16/3 16.4
function, 15/35
long-term sales programmes, 15/35–8
market research, 15/36
quality, 17/4
short-term sales programmes, 15/38
statistical forcasting, 15/36
Martensite steel, 1/23–4
Martensitic stainless steel, 1/12, 1/36–7
Martin Marietta alloys, 1/84
Mashy state,
composite sheet manufacture, 4/128–9
forging particle-reinforced metals, 4/128–9
metal extrusion, 4/127–8
metal rolling, 4/128
Master production schedules (MPS), 14/5–6
Materials acquisition,
classical selection procedure,
component requirements, 1/6
environmental resistance, 1/7
functional analysis, 1/6
manufacture and cost, 1/7
mechanical properties required, 1/7
physical properties required, 1/7
property requirements, 1/6
requirement check-list, 1/6–7
wear resistance, 1/7
comparative selection procedure, 1/8
imitative selection procedure, 1/8
information sources, 1/8
materials requirements planning (MRP), 14/5–7, 14/7–8, 15/50
materials selection techniques, 1/6
selection computerisation, 1/8–10
PERITUS knowledge-based system, 1/9–10
Mathematical modelling,
metal forming processes,
Coulomb friction correction, 4/21, 4/22
equilibrium technique, 4/19–20
finite-element process, 4/23
free surface, 4/21
frictionless surface, 4/21
Hencky's equation, 4/22
mean yield stress, 4/22
power of deformation, 4/22
rolling, 4/54
slip-line analysis, 4/19–20
slip-line field, 4/21
upper-bound method, 4/22
visioplastic technique, 4/22–3
Maximum time to repair equipment, 18/6
MCAD (mechanical CAD), 14/4
Mean time between failures (MTBF), 18/6
Mean time to repair equipment, 18/6
Measles charts, 17/20, 17/21
Measurement/measuring instruments,
see also Radiography; Transducers; Ultrasonics

accuracy, 17/33
air gauges, 17/41
alignment principle, 17/33–4
angular measurement, 17/35–6
bevel protractors, 17/35
circular measurement, 17/39–40
comparators, 17/35
computer vision analysis, 17/43
coordinate measurement machines, 17/42–3
dial gauges, 17/34
dial test indicators, 17/34
diameter gauges, 17/37–8
feeler gauges, 17/34
gauge blocks, 17/35
gauge material, 17/38
hole depth gauges, 17/37
inspection probes, 17/41–2
laser use, 17/43
length bars, 17/35
limit gauges, 17/36–8
linear measurement, 17/34–5
micrometers, 17/34, 17/35
National Physics Laboratory (NPL), 17/33
optical systems, 17/43
reactivity, 17/33
repeatability, 17/33
roundness measurement, 17/39–40
screw tread gauges, 17/37
sensitivity, 17/33
sine bars and tables, 17/35–6
slip blocks, 17/35
spirit levels, 17/36
steel rules, 17/34
stylus measurement, 17/39–40
surface plates, 17/35
taper limit gauges, 17/37
Taylor's theorem of gauging, 17/36–7
texture of surfaces, 17/38–9
threads,
form (flenk angle), 17/44
internal, 17/46
large, 17/46
major diameter, 17/43–4
with microscopes, 17/44–5
minor diameter, 17/44
pitch, 17/45
with projectors, 17/44
simple effective diameter, 17/45–6
vernier calipers, 17/34
virtual effective diameter, 17/46
Mechanical CAD (MCAD), 14/4
Media access control (MAC), 13/19
Media, network, 13/19
Melamine-formaldehyde (MF) plastics, 2/25
Memory, microprocessors,
EPROM (erasable read only memory), 11/38
RAM (ramdom access memory), 11/38
ROM (read only memory), 11/38
Mesh analysis, 10/4
Metabel explosive for high-energy-rate metal forming, 4/115
Metal,
casting *see* Casting; Die casting; Sand casting
crystalline structure of, 1/15
drawing *see* Cold drawing
extrusion *see* Extrusion
fibre reinforced *see* Composites
flow *see* Plastic flow; Plasticity
forging *see* Forging
forming *see* Forming metal
machining/cutting *see* Machining
sheet forming *see* Sheet metal forming
Metal oxide semiconductor field effect transistor (MOSFET), 10/22
Mice, computer, 13/12–13
Microcomputers *see* Computers
Microcontrollers,
see also Direct digital control
architecture, 11/53
Mitsubishi series 740 8–bit CMOS, 11/53
Motorola MC 68705R3, 11/53–4
Rockwell R6500/1, 11/54
Micrometers, 17/34, 17/35
Microprocessors,
see also Computers; Microcontrollers
ASCII code, 11/36
binary system, 11/36
bus structure, 11/37–8
EPROM, 11/38
Gray code, 11/36–7
hexadecimal system, 11/36
IEEE (Hewlett-Packard) interface, 11/41
industrial controllers, 11/54
input/output,
dedicated port addressed, 11/39
direct memory access (DMA), 11/39
memory mapped, 11/39
structure, 11/38–9
support chips, 11/39
memory maps, 11/39–40
RAM, 11/38
ROM, 11/38
RS232C, RS422, RS423 serial interfaces, 11/40–1
single-chip microcontrollers (SCMs), 11/52–3
software, 11/54–5
system architecture, 11/37
Milling machines, 5/4–5, 5/7, 5/9
Mills *see* Rolling metal
Minicomputers *see* Computers
MIS (manufacturing information systems), 14/4
Modelling *see* Mathematical modelling
Modems, data transmission, 13/20
MODUS management recording system, 15/12
Moire fringe effect, 11/24
Molybdenum,
characteristics and applications, 1/93
cold drawing, 4/97
Monels, 1/81, 1/83
explosive forging, 4/125–6
flow forming, 4/52
Monitoring, production *see* Management, systems theory
Monitors, CADCAM, 14/25–6
MOSFET (metal oxide semiconductor field effect transistor), 10/22
Motion study, 15/18
Motors, a.c. *see* Induction motors; Synchronous motors
Motors, d.c.,
see also Stepper motors
armature voltage speed control, 10/50
brushless, 10/59, 12/5–6
chopper speed control, 10/51
compound-wound, 10/49
disk rotor, 12/5
efficiency, 10/51
field regulator speed control, 10/50
microprocesser speed control, 11/57–8
overload cut-out, 10/49
permanent magnet, 10/57, 12/5
series-wound, 10/49
shunt-wound, 10/48
speed control, 10/50–1
starting, 10/49–50
stepping,
hybrid, 12/6
permanent magnet, 12/6
variable reluctance, 12/6
torque, 10/48
universal a.c./d.c., 10/57
Ward Leornard speed control, 10/50–1
Moulds/moulding,
see also Dies; Extrusion
foam, stiffness of, 2/44
metal,
core sand, 3/8
die casting, 3/14
facing sand, 3/8
investment castings, 3/17, 3/18–19
liquid sand, 3/8
loose sand, 3/8
moulding sands, 3/7–9
mould making, 3/5–8
permanent moulds, 3/7
semipermanent moulds, 3/7
shell, 3/19–20
single use (expendable) moulds, 3/6, 3/7–9
system sand, 3/8
plastics and rubbers,
blow, 2/33–5
compression, 2/28
injection, 2/27, 2/29, 14/80–1
rotational, 2/37
transfer, 2/28–9
powders, 3/21–3
Mountings, rubber, 2/38–40
MPS (master production schedules), 14/56
MRP II (manufacturing-resource planning), 14/7–8
MRP (materials requirements planning), 14/5–7, 14/7–8, 15/50
MTBF (mean time between failures), 18/6
Multidisciplinary team working, 16/8
Multiplexing, computers, 11/48–50
Mumetal, hysteresis in, 10/10

National Physics Laboratory (NPL), 17/33
Natural rubber, 2/13
NC (numerical control) *see* Computer numerical control; Direct numerical control
NE (numerical engineering), 14/27–8
NetBIOS, networks, 13/19
Network layer, OSI, 13/19
Networks, computer, 13/17–20
Arpanat, 13/22
data transmission standards, 13/21
Ethernet, 13/23
Euronet, 13/22–3
glossary of terms, 13/18–20
LANs (local area networks), 13/22
modems, 13/20
Prestel, 13/22
protocols, 13/21–2
Telenet, 13/22
topologies, 13/21
transmission media, 13/20–1
WANs (wide area networks), 13/22
NFS (network file system), 13/19
Waspaloy, 1/86
Nickel/nickel alloys,
Astroloy, 1/86
composition table, 1/85
corrosion resistant alloys, 1/81–2, 1/83
creep resistant, 1/82
directionally cast eutectics, 1/86
directional solidification, 1/84, 1/86

Nickel/nickle alloys, *continued*
electrical alloys, 1/87, 1/88
expansion controlled, 1/87, 1/89
50/50 nickel chrome, 1/82
for gas turbines,
casing materials, 1/87
cast blades, 1/83, 1/85
composite blades, 1/84, 1/85
forged blades, 1/82
turbine-disc alloys, 1/85, 1/86
wrought, 1/85
general characteristics, 1/81
grain boundary structure, 1/84
hard facing materials, 1/87
Hastelloys, 1/81
high temperature, 1/82
hot extrusion, 4/78
Iliums, 1/81, 1/83
Inconels, 1/81–2, 1/83
magnetic alloys, 1/87, 1/88
Martin Marietta alloys, 1/84
mechanically alloyed powder, 1/85
Monels, 1/81, 1/83
Nimonic, 1/82, 1/84, 1/86
nomenclature, 1/81
single crystal casting, 1/84
standards, 1/81
tungsten addition, 1/84
Nimonic alloys, 1/82, 1/84, 1/86
laser-beam machining, 6/7
Niobium, characteristics, 1/93
Nitrile rubbers, 2/14
Nodal analysis, 10/4
Noise, amplifiers, 10/24
Non-magnetic fittings, copper alloys for, 1/62
Norton's theorem, 10/4
Noryl (modified PPO), 2/25
NPL (National Physics Laboratory), 17/33
Nuclear metals,
beryllium, 1/94
canning metals, 1/94
fissile metals, 1/94
zirconium, 1/94
Numerical control (NC) *see* Computer numerical control; Direct numerical control
Numerical engineering (NE), 14/27–8
Nuts,
see also Threads
free spinning, 9/5
lock, 9/5
stiff, 9/7, 9/8
Nylon,
for engineering, 2/24, 2/27
water-jet machining, 6/5
Nyquist stability criteria, 11/10–12

Ohms law, 10/3, 10/4
Oil pipelines, steel for, 1/26
Oil tank heaters, copper alloys for, 1/62
Open-die forging *see* Forging
Open loop control, 11/4
Open system interconnection (OSI), 13/19
application layer, 13/18
with PLCs, 11/64
Operating cycle, 18/6
Operating systems, (computers), 13/5–6
1100 Executive, 13/7
DOS, 13/7
MCP, 13/7
OS/360, 13/7
TOPS, 13/7
Unix, 13/7
Optical character recognition, 14/86
Optimum production technology (OPT), 15/50
philosophy, 14/8
scheduling, 14/8–9
Opto-isolators, 11/42
Orbital forging *see* Forging
Ordering *see* Production control, ordering
OSI *see* Open system interconnection
Osmium, 1/94
Oxygen argon lancing, 1/13
Oxygen blown steel, 1/13
Oxygen converter process (steel), 3/4
Oxygen lance heads, copper alloys for, 1/62

Packet, network, 13/19
Pallets,
carrier, 14/63
fixture, 14/63
handling methods, 14/63
work change stations, 14/63
Paper, water-jet machining, 6/5
Paper tape readers, CNC systems, 14/50–1
Parallel engineering, 16/8–10
Paraorientated aromatic polyamide fibres, 1/96
Pareto curves, 17/18, 17/19
Patterns, metal casting,
gated patterns, 3/4–5
investment casting, 3/17
loose with core prints, 3/4–5
loose patterns, 3/4
PCBs *see* Printed circuit boards
PCs *see* Personal computers
Peak inverse voltage voltage (PIV), rectifiers, 10/18
Pearlite *see* Steel
Peer-to-peer resource sharing, networks, 13/19
Pemserter fastener, 9/3
Penetrant testing,
development, 17/55
postcleaning, 17/55
precleaning, 17/55
PERITUS material selection system, 1/9–10
Permalloy, hysteresis in, 10/10
Permeability/permeance, 10/7
Permender, 1/92
Permittivity, 1/5
Personal computers (PCs), 11/52
advantages/disadvantages over PLCs, 11/60
for CADCAM, 14/24–5
Eurocards, 11/58
IEE–488 (GPIB) interface, 11/58–9
for instrumentation and control, 11/58–9
with shop floor data collection, 14/10
Personnel,
employee job satisfaction, 15/55
employee morale, 15/55
as management function, 15/55
PFA *see* Production flow analysis
Phasor diagrams, 10/14
Phenolic adhesives, 9/28
Phenolic plastics, 2/25
Photo-plotters, CADCAM, 14/24
Photosensors *see* Sensors
Physical layer, OSI, 13/19
PID (proportional, integral and derivative) control, 11/14–15, 11/54
digital control algorithm, 11/56–7
Piercing,
geometry factors, 4/19
strain and strain-rate relationships, 4/16–17
Piezoelectric effect transducers, 11/28–9, 17/52
Pilger mill, rotary tube forging, 4/66–9
Planing machines, 5/3
Planning, quality, 17/4
Plant,
see also Maintenance
procurement, 18/12
replacement, 18/12
selection, 18/11
Plasma arc,
applications, 6/8
cutting, 6/8
machining, 6/8
welding, 6/8
Plastic flow of metal,
see also Plasticity
Bauschinger effect, 4/3
compressive stress systems, 4/6
crystalline structure slip, 4/3
elastic recovery, 4/3
extrusion, 4/10
high energy rate processes, 4/10–11
interfacial slip, 4/3
operational strain rate, 4/6
operational-stress system, 4/5–6
operational temperatures, 4/4–5
process classification, 4/4–6
rolling,
longitudinal, 4/6, 4/8
oblique, 4/6, 4/8, 4/9
profiling and shaping, 4/8, 4/10
sheet and plate, 4/5, 4/8
tube, 4/5–7
shape-effect criterion, 4/5
stress classification, 4/7
tensile compressive systems, 4/5–6
tensile stress systems, 4/6
Plasticisers, 2/6
Plasticity,
see also Stress-strain relationships
basic law, 4/13
deformation zones, 4/13–15
geometry factor, 4/18
Levy-Mises rule of flow, 4/19
Poisson's ratio, 4/19
redundancy, 4/18, 4/20
simulation of metal flow, 4/23–7
with plasticine, 4/24–5
with wax, 4/24–6
strain redundancy factor, 4/18
strain and strain-rate relationships, 4/15–17
yielding and yielding criteria, 4/17–18
Huber-Mises criterion, 4/18
Tresca criterion, 4/17–18
Plastics,
see also Polymers
ageing, 2/43
anistropy, 2/43
bulk compression, 2/43
common names and abbreviations for, 2/23
component design, 2/43–5
creep, 2/19–20, 2/43–4
design with, 2/43–5
fibre-reinforced laminates, 2/36–7
flame retardancy, 2/17–18
foam mouldings, 2/44
processing, 2/26–37
calendering, 2/35
flow path ratio, 2/43
frictioning, 2/35
in molten shape, 2/26–8
moulding, 2/27, 2/28–31, 2/33–5
setting, 2/26
shaping, 2/26

thermoforming, 2/35–6
shrinkage, 2/43
softening point, 2/16
solvent resistance, 2/20, 2/20–1
stiffness, foam mouldings, 2/44
stress concentrations, 2/43
thermal stability, 2/16–17
thermosetting, 2/3
threaded inserts, 9/7
toughness, 2/18–19
ultrasonic welding, 9/7
water-jet machining, 6/5
Plate, rolling *see* Rolling metal
Plates, steel, 1/26
Platinum, applications, 1/94
PLCs *see* Programmable logic controllers
PLDs (programmable logic devices), 10/43
Plotters,
electrostatic, 14/27
pen, 14/27
thermal, 14/27
Plug mill system, metal rolling, 4/43
Pneumatic systems,
CNC machines, 14/32
PLC control, 11/62–3
Pneumatic valves and actuators, 12/9–12
Poisson cumalitive probability tables, 17/59
Poisson's ratio, 1/3, 2/7, 4/19
Polishing,
barrelling,
compounds, 8/ 8.33
equipment, 8/ 8.33
medium, 8/ 8.33
polishing/finishing, 8/ 8.33
micropolishing, 8/ 8.33
processes, 8/33
Polyacetals, 2/24, 2/27
Polyamides, 2/24
Polyblends, 2/6
Polybutadiene rubber, 2/14
Polybutylene terephthalate (PBT), 2/25
Polycarbonates, 2/24, 2/27
Polychloroprene rubbers, 2/14
Polyester resins, thermosetting, 2/25–6
Polyether ether ketone (PEEK), 2/25, 2/27
Polyether ketone (PEK), 2/25
Polyether sulphones (PES), 2/24, 2/27
Polyethylene, 2/3, 2/22
Polyethylene terephthalate (PET), 2/25, 2/27
Polyisoprene rubber, 2/14–15
Polymerisation, 2/3
Polymerisation casting, 2/36
Polymers,
see also Composites, polymer; Plastics; Rubbers
accelerators, 2/5–6
additives, 2/5–6
blends, 2/6
elasticity in, 2/4–5
fillers, 2/6
flow properties, 2/26
LOI (limiting oxygen index), 2/17–18
plasticisers, 2/6
processing, 2/26–37
heat required for, 2/27
in molten state, 2/26–8
setting, 2/26
shaping, 2/26
temperatures for, 2/27
to metal adhesion, 2/26
Polymethyl methacrylate, 2/25, 2/27
Polymides, 2/25
Polynorbornenes, 2/15
Polyolefins, 1/101
Polyphenylene sulphide (PPS), 2/24–5
Polypropylene, 2/22, 2/27
Polystyrene, 1/101, 2/22–4, 2/27
Polysulphides, 2/15
Polysulphones (PSU), 2/24, 2/27
Polytetrafluoroethylene (PTFE), 2/25
Polytetramethylene terephthalate (PTMT), 2/25
Polyurethane adhesives, 9/28
Polyurethane foams, 2/37
Polyvinyl chloride, 1/101
Polyvinyl chloride (PVC), 2/22
Porcelain *see* Ceramics
Portholes, copper alloys for, 1/62
Potential difference, 10/4
Powder compaction, high-energy-rate techniques, 4/121–2
Powder metallurgy, 3/21
Power,
electrical, 10/4
power factor, 10/15
three-phase circuits, 10/52–3
Power of deformation, modelling metal formation, 4/22
Power supplies, dc from ac, 10/27–9
reservoir capacitors, 10/28
ripple low pass filters, 10/28–9
PPVC (plasticised PVC), 2/22
Precipitation hardening stainless steel *see* Stainless steel
Presentation layer, OSI, 13/19
Presses, CNC sheet metal punch, 14/44–5
Pressure die casting *see* Die casting
Pressure vessels, copper alloys for, 1/61
Pressure vessel steels, 1/26–7
Prestel network, 13/22
Printed circuit boards (PCBs),
ATE testing, 7/12
CADCAM,
application, 14/22–4
circuit layouts, 14/24
circuit simulation, 14/24
component/symbol database, 14/24
engineering analysis, 14/24
NC drilling machine output, 14/24
net lists, 14/24
photo-plotter output, 14/ 14.24
routing, 14/24
schematic design capture, 14/24
cleaning, 7/12
layout, 7/12–13
manufacturing defects, 7/13
mixed through hole and surface mount assembly, 7/8
soldering,
conductive belt, 7/12
infra-red, 7/11
laser reflow, 7/11–12
vapour phase reflow, 7/11
wave soldering, 7/10
surface mount assembly,
double sided, 7/7
single sided, 7/7
surface mount components,
active, 7/4–5
attachment to boards, 7/9
packaging, 7/6–7
passive, 7/5–6
solder pastes, 7/10
surface mount placement machines, 7/12
surface mount technology, 7/4
introduction, 7/13
through hole component assembly,
automated, 7/4
manual, 7/3
Printers,
dot-matrix, 13/10, 14/27
laser, 13/10, 14/27
Print server, 13/19
PRN (process route numbers), 15/26
Probes, measurement *see* Measurement/measuring instruments
Process route numbers (PRN), 15/26
Procurement, plant *see* Plant
Product design, 15/14–15
design modifications, 15/14
quality control, 15/14–15
Production,
see also Group technology; Production control; Production planning
batch production methods, 15/34
combined classification, 15/5, 15/6
continuous line flow, 15/19–20
factory flow analysis (FFA), 15/23, 15/26, 15/27
group technology, 15/18, 15/20–2, 16/4–5
machine paced assembly lines, 15/34
market type classification,
batch-production, 15/5
continuous-production, 15/5
jobbing, 15/5
large-contract, 15/5
materials conversion classification,
explosive industries, 15/5
implosive industries, 15/4–5
process industries, 15/4, 15/5
square industries, 15/5
process route numbers (PRN), 15/26
production flow analysis (PFA), 15/23, 15/26
traditional assembly organisation, 15/32, 15/34
workshop level organisation, 15/18–19
Production control,
annual programmes, 15/42
by industry type, 15/40
component scheduling, 15/49
computer integrated manufacturing (CIM), 15/50
dispatching,
Gantt chart scheduling, 15/50
inspection, 15/50
operation scheduling, 15/50–1
period batch control (PBC), 15/50
as production control function, 15/50–2
economic batch quantity (EBQ), 15/44–6
explosive industries, 15/40
flexible programming, 15/42–3
implosive industries, 15/40
inventory control, 15/40
just-in-time system, 15/46, 15/49
kitting out for assembly, 15/51
loading activity, 15/40
materials requirement planning (MRP), 15/50
optimum production technology (OPT), 15/50
ordering, 15/38, 15/40
base stock control method, 15/46
flow control methods, 15/48
kanban system, 15/46
multi-cycle, 15/43–6
re-order level methods, 15/46–7
requirement cover methods, 15/47–8
single cycle, 15/43–6
stock and allocation control, 15/47
stock-base, 15/43, 15/46–8
stock and provision control, 15/47
stock ratio optimisation (STROP), 15/47–8
period batch control (PBC), 15/49–50
process industries, 15/40
programme reconciliation, 15/43

Production control, *continued*
  programming,
    introduction, 15/38, 15/40
    levels of, 15/36
  progressing, 15/40
  scheduling, 15/40
  unsalable stock accumulation, 15/42
Production flow analysis (PFA), 15/23
  see also Group technology
Production planning,
  just-in-time (JIT) production, 15/15
  main decision types, 15/15
  make or buy decisions, 15/15–16
  as management function, 15/15
  material form choice, 15/16
  planning specifications, 15/15–16
  processing route choice, 15/16–17
  production control policies, 15/15
  tool requirement planning,
    costs, 15/17
    quality needed, 15/17
    set-up times, 15/17
    speed requirement, 15/17
  work study techniques, 15/18
Productivity, welding, 9/26–7
Product specifications,
  drawings, 15/14
  material specifications, 15/14
  parts lists, 15/14
Programmable logic controllers (PLCs),
  advantages/disadvantages over PCs, 11/60
  applications, 11/59
  categories of, 11/59
  hydraulic system control, 11/62–3
  ladder logic programming, 11/60–2
    counters, 11/62
    input, 11/61
    output, 11/61
    timers, 11/61
  networking, 11/63–4
  pneumatic system control, 11/62–3
  safety aspects, 11/63
Programmable logic devices (PLDs), 10/43
Programming,
  see also Software
  CNC systems,
    APT, 14/56–7
    canned (fixed) cycles, 14/56
    conversational, 14/57–8
    graphical, 14/57
    macros, 14/55
    manual, 14/51, 14/54–6
    methods and types, 14/51
    subroutines, 14/55–6
  robots, 11/67–8, 14/77–9
Programs *see* Software
Propellors, copper alloys for, 1/62
Protocol, definition, 13/19
Proximity sensing, 11/29
PTFE (polytetrafluoroethylene), 2/25
Pulse-width modulation (PWM), 11/44–5
Pump components, copper alloys for, 1/61
Purchasing,
  batch buying, 15/52–3
  call-off buying, 15/53
  schedule buying, 15/53
  single sourcing, 15/53–4
  supplier assessment, 15/54
  supplier selection, 15/53–4
PVC (polyvinyl chloride), 2/22
  spreading and dipping, 2/37
PWM (pulse-width modulation), 11/44–5
Pyrolysis, ceramics, 1/115

Q factor, electrical, 10/16
QIPs (quality improvement programs), 16/7
Quality,
  see also British Standards Institution; Quality assurance; Quality control; Vendor assessment
  adaptive control, 17/16
  assembly analysis, 17/10
  a competitive weapon, 17/3–5
  conceptual design evaluation, 17/8
  costs,
    appraisal, 17/5
    external failure, 17/5
    internal failure, 17/5
    measurement of, 17/5–6
    prevention, 17/5
  definition, 17/3
  design modifications,
    concessions, 17/7
    control of, 17/7
    production permits, 17/7
  design, role of, 17/7
  design techniques for, 17/8–10
  failure mode, effect and criticality analysis, 17/9–10
  grade, 17/3
  inspections,
    first-off, 17/15
    last-off, 17/15–16
    post-process, 17/16
    preprocess, 17/15
  law aspects,
    Consumer Protection Act, 17/6
    contracts, 17/6
    Health and Safety at Work Act, 17/6
    Trades Description Act, 17/6
    vicarious liability, 17/6
  material control, 17/13, 17/15
  problem identification,
    correlation charts, 17/20
    Cusum charts, 17/18–19, 17/20
    Fishbone diagrams, 17/18, 17/19
    measles charts, 17/20, 17/21
    Pareto curves, 17/18, 17/19
  process control, 17/15–17
  process selection, 17/8
  process variability,
    dispersion patterns, 17/17
    drift patterns, 17/17–18
  quality improvement programmes (QIPs), 16/7
  quality level, 17/3
  reliability assessment, 17/9
  specification check list, 17/8
  specifications, 17/7
Quality assurance,
  see also Vendor assessment
  design control, 17/4
  information and equipment duties, 17/7
  management objectives, 17/4
  manufacturing control, 17/4
  marketing, 17/4
  material control, 17/4
  planning, 17/4
  process control engineering duties, 17/7
  programme, 17/4
  purchasing control, 17/4
  quality control, 17/4
  quality engineering duties, 17/7
  servicing, 17/4–5
Quality control,
  attribute control charts,
    description, 17/30–1
    number defective type, 17/31
    percentage defective type, 17/31–2
  double sampling plans, 17/58
  as management function, 15/14–15
  normal distribution, 17/26
  Poisson cumalitive probability tables, 17/59
  sample size code letters, 17/56
  single sampling plans, 17/57
  variable control charts,
    data collection for, 17/27
    description, 17/26–7
    limit setting, 17/27–9
    operation, 17/29
    relative precision index (RPI), 17/28–9
Quenching, heat treatment, 1/29

Radiators, car, copper alloys for, 1/61
Radiofrequency identification, 14/86–7
Radiography,
  flaw detection, 17/48
  gamma radiation, 17/47–8
  inverse square law, 17/46
  radiation intensity, 17/46
  radioactive decay, 17/47
  radiographs, 17/48
  safety, 17/49–50
  weld radiographs, 17/48–9
  X radiation, 17/46
  X-ray equipment, 17/46–7
Rail guided vehicles (RGVs), 14/59, 14/63, 14/65
Rake, cutting tools, 5/6–10, 5/13, 5/14, 5/15
Random access memory (RAM), 11/38, 13/8–9
Raster display monitors, CADCAM, 14/25
Reaction bonding, ceramics, 1/115
Reactive power, 10/15
Read only memory (ROM), 11/38, 13/9
Records, asset register, 18/4
Rectifiers/rectification, 10/17–18
Redirector, software, 13/19
Reducing mills, 4/43
Refraction, sound, 17/50–1
Refractive index, light, 1/5
Registered firm of assessed capability mark, BSI, 17/12
Reinforced concrete *see* Concrete, reinforced
Reinforced thermoplastics (RTPs), 1/101
  see also Composites
Reinforcing fibres *see* Composites
Relative permittivity, 1/5
Relative precision index (RPI), 17/28–9
Reliability,
  assessment, 17/9
  equipment, 18/6
Reluctance, magnetic, 10/7
Remanence, magnetic, 10/10
Remote file service, 13/19
Remotely operated vehicles (ROVs), 11/52
Repeaters,
  buffered, 13/18
  cable, 13/19
Residual stresses, in metal forming tools, 4/33–4
Resistance/resistivity, 1/5, 10/3
Resonance, electrical, 10/16
Return on investment (ROI), 18/10
RGVs (rail guided vehicles), 14/59, 14/63, 14/65
Rhodium, 1/94
RH (Ruhr-Stahl Heraes) steel process, 1/13
Risers, sand casting, 3/12–14
Rivets, 9/8, 9/9–10
Robots,
  accuracy, 14/83
  applications, 11/64–5, 14/79–82
    arc welding, 14/81

assembly, 14/81
injection moulding, 14/80–1
spot welding, 14/81
surface coating, 14/81–2
at turning centres, 14/43–4
British Robot Association, 14/73
computational hardware, 11/68
configuration, 14/83
construction, 14/74–7
control interfaces, 11/50–2
control unit, 14/74
costs, 14/83
definition, industrial robotics, 14/73–4
electrically powered, 14/75
ergonomic aspects, 19/4–5
geometric configuration classification, 14/75, 14/76
geometry, 11/65–7, 14/74–6
hydraulically powered, 14/75, 14/77
loads, 14/83
path control, 11/68, 14/75, 14/77
pneumatically powered, 14/77
position control, 11/66
precision, 14/83
programming,
AML (IBM), 11/68
languages,14/78–9
lead-through, 11/68
methods, 14/83
off-line, 14/77
on-line, 14/77–8
requirements from, 11/67
simulation, 14/79
teach pendants, 11/68
VAL (Unimation), 11/68
walk-through, 11/68
PUMA (Unimate) hardware, 11/68
reasons for use, 14/74
safety,
accidents, 14/83–5
hardware integrity, 14/83
software integrity, 14/83
SCARA type, 11/66–7, 14/75, 14/81, 14/82
selection of, 14/82–5
speed considerations, 14/82–3
velocity control, 11/66
Rockwell hardness, 1/3–4
Rod drawing *see* Cold drawing
ROI *see* Return on investment
Rolling metal,
see also Tube rolling
basic theory, 4/54–6
carbon steels, 1/18
classification of processes, 4/35–8
cold operations, 1/18, 4/41
theory, 4/54
finish of surface, 4/38
H-type mills, 4/42
I-beams, 4/42
longitudinal, 4/6, 4/8, 4/42, 4/43–5
machine parts, 4/53–4
Manfred-Weiss process, 4/45
Mannesmann cycle, 4/45
mashy state, 4/128
mathematical modelling, 4/54
mill classes,
cogging, 4/35
plate, strip and sheet, 4/35
special, 4/35
tube, 4/35
mill layouts, 4/40
mill trains, 1/20
mill types,
double duo, 4/36
four high, 4/36
HC mill, 4/36
Sendzimir twelve high, 4/37, 4/41
six high, 4/37
Taylor, 4/37
three high, 4/36
twenty high, 4/37
two high, 4/36
oblique, 4/6, 4/8, 4/9
plastic flow,
longitudinal, 4/6, 4/8
oblique, 4/6, 4/8, 4/9
sheet and plate, 4/5, 4/8
tube, 4/5–7
plate, 4/40–2
profiling and shaping, 4/8, 4/10
recrystallisation temperature, 4/38
roll design, 4/38–40
roll profiles,
box pass, 4/39, 4/40
circular groves, 4/39, 4/40
diamond passes, 4/39, 4/40
oval groves, 4/39, 4/40
special, 4/39, 4/40
roll speed, 4/38
roll systems,
horizontal-vertical, 4/37
universal slabbing, 4/37
universal structural, 4/37
vertical, 4/37
vertical-horizontal, 4/37
Senzimir cluster mill, 1/20
sheet, 4/40–2
sheet and plate, 4/5, 4/8
sheet roll bending, 4/54
sizing mills, 4/43
strain hardening, 4/38
stretch reducing mills, 4/44–5
strip, 4/40–2
structural sections, 4/42
temperature of processing, 4/38
T-section mills, 4/42
ROM (read only memory), 11/38, 13/9
Roofing, copper alloys for, 1/62
Rotaform, orbital forging process, 4/69
Rotary forging,
bar stock, 4/66
tubes, 4/66–9
Router, network, 13/19
Routh-Hurwitz stability criterion, 11/12
ROVs (remotely operated vehicles), 11/52
RPI (relative precision index), 17/28–9
RTPs *see* Reinforced thermoplastics
Rubbers,
see also Polymers
additives, 2/5–6
ASTM D2000 classification, 2/10–13
bridge bearings, 2/38
bushes, 2/38
component design with, 2/37–42
creep and set, 2/10
damping with, 2/38
design with, 2/37–42
exotic, 2/15
hardness, 2/7
incompressibility of, 2/37
laser-beam machining, 6/7
liquid processing, 2/15
metal bonding, 2/38
modulus, 2/7
mountings, 2/38
processing, 2/26–37
calendering, 2/35
frictioning, 2/35
in molten state, 2/26–8
moulding, 2/28–31, 2/33–5
setting, 2/26
shaping, 2/26
thermoforming, 2/35–6
properties, non-Rubbery, 2/10
properties, rubbery materials summary, 2/12
resilience, 2/9
seals, 2/41
speciality, 2/15
special purpose, 2/14–15
stress-strain properties, 2/6–9
survey of types, 2/11–15
thermoplastic, 2/5, 2/15
tyre, 2/13–14
u-rings, 2/41–2
vibration isolation, 2/41–2
vulcanised, 2/4–5
water-jet machining, 6/5
Ruhr-Stahl Heraes (RH) steel process, 1/13

SAA (system application architecture), 13/19
Safety,
adhesives, 9/33
electrical, 10/62–3
circuit breakers, 10/63
double insulation, 10/63
earthing, 10/63
earth leakage protection, 10/63
electric burn, 10/62
electric shock, 10/62
fuses, 10/63
insulation, 10/62–3
isolation, 10/63
low voltage supplies, 10/63
protection, 10/62–3
rescue, 10/62
with PLCs, 11/63
radiography, 17/49–50
with robots, 14/83–4
Safety mark, BSI, 17/12
Sales *see* Marketing
Salt bath furnace, 4/79
Sampling, product acceptance,
BS 6001 procedures and tables,
acceptable quality level (AQL), 17/23
average outgoing quality level (AOQL), 17/24–5
code letter, 17/22
inspection class, 17/23
inspection level, 17/22
sample plan detail, 17/23
sample selection, 17/23
testing curtailment, 17/23
double sampling, 17/21–2
inspection frequency, 17/20–1
multiple sampling, 17/22
non-standard quality plans, 17/25–6
operating-characteristic curve,
average outgoing quality level (AOQL), 17/24–5
construction, 17/24
sampling plan criteria effects, 17/24
use, 17/23–4
plans for normal inspection,
double, 17/58
single, 17/57
sample size code letter chart, 17/56
sampling plan, 17/22
sequential sampling, 17/22
single sampling, 17/21
Sand casting, 3/4–14
see also Moulds/moulding; Patterns
contraction in, 3/5–6
gating system, 3/9–12
moulding, 3/5–9

Sand casting, *continued*
pattern making, 3/4–5
pouring rates/times, 3/10–11
risers, 3/12–14
shell moulding, 3/19–20
Sand/quartz, grinding abrasive, 8/4
SATE FMS system, 14/66–7
Schedule buying, 15/53
SCMs (single chip microcontrollers) *see* Microcontrollers
Screws,
see also Threads
copper alloys for, 1/62
head forms, 9/4
self-tapping, 9/6, 9/7
thread-forming, 9/6
SCRs (Silicon controlled rectifiers), 11/42
see also Thyristors
Seals, rubber, 2/41
Seawater tubes, copper alloys for, 1/62
Secretarial activity,
computer use, 15/55
as management function, 15/55–6
Selas radiant heat type furnace, 4/79
Selection, plant *see* Plant
Self bonding, ceramics, 1/115
Semiconductors,
covalent bonds, 10/16
n-type, 10/16
photosensors, 12/15
p-n junction diode, 10/17
p-type, 10/16
storage devices, 13/8–9
Sendzimir twelve high mill, 4/37, 4/41
Sensors,
see also Temperature measurement; Transducers
capacitive, 12/14–15
inductive, 12/14
photosensors,
cadmium sulphide, 12/15
photosensitive semiconductors, 12/15
photovoltaic, 12/15
pneumatic,
back pressure, 12/15
interruptible jet, 12/15
Senzimir cluster mill, 1/20
Server, computer, 13/19
Server message block (SMV), networks, 13/19
Servicing, quality, 17/4–5
Servodrives, CNC machines, 14/30–1
Servomechanisms, 11/3
Servomotor control amplifiers,
four-quadrant control, 12/7–8
pulse-width modulation, 12/7–8
Servomotors *see* Motors d.c.
Session layer, OSI, 13/19
SFDC *see* Shop floor data collection
SG abrasives, 8/17–18
Shaft encoders, 11/23
Shannon's sampling theorem, 11/56
Shaping machines, 5/3
Shearing metal, 4/107–8
Shear modulus, 1/3, 2/7–8
Shear power, modelling metal formation, 4/22
Sheet metal forming, 4/4
see also Rolling
bending, 4/108–9
minimum permissible radius, 4/108
overbending, 4/108
continuous tandem rolling mill, 4/106–7
deep drawing, 4/111–13
defects, 4/110
domestic applications, 4/107
fine blanking, 4/108, 4/109
finishes, 4/107
flanging, 4/110
hydro and hydromechanical, 4/111
industrial applications, 4/107
rolling, 4/40–2
shearing, 4/107–8, 4/109
sheet roll bending, 4/54
spinning, 4/110
stretch forming, 4/110–11
Sheet metal punch presses, CNC, 14/44–5
Shell (investment) moulding, 3/19–20
Shop floor data collection (SFDC),
bar code readers, 14/10–11
central computer systems, 14/10–11
data collection methods, 14/9
information types collected, 14/9
machine data units, 14/10
shop floor terminals, 14/10
use of PCs, 14/10
Shrinkage, plastics, 2/43
Sialons, 1/112
Siemens-Martin steel manufacturing process, 3/4
Sigal-to-noise ratio, amplifiers, 10/24
Silicon, in integrated circuits, 1/94
Silicon carbide,
applications, 1/112
commercial products, 1/111–12
fibres, 1/96
grinding abrasive, 8/4
properties and selection, 1/116
Silicon controlled rectifiers (SCRs), 11/42
see also Thyristor
Silicone adhesives, 9/28
Silicone rubbers, 2/15
Silicon nitride,
ceramic tools, 5/11
manufacture, 1/112
properties and selection, 1/116
Silver, applications, 1/93–4
Simulation,
manufacture, computer based, 14/11–12
metal flow, 4/23–7
for forging with computer, 4/72
with plasticine, 4/24–5
with wax, 4/24–6
Simultaneous engineering, 16/8–10
Sine bars and tables, 17/35–6
Single-chip microcontrollers (SCMs) *see* Microcontrollers
Sintered carbide cutting materials, 5/10–11
Sintered ceramic cutting tools, 5/11
Sintered cutting inserts, 5/3
Sintering, 3/21–3
ceramics, 1/115
compacting, 3/22
temperarure effects, 3/22
Slab technique, metal forming modelling, 4/19–20
Slide movement, CNC machines,
position measurement, 14/32
rack and pinion, 14/32
recirculating ball and leadscrew, 14/32
Slip blocks, 17/35
Slip-line analysis, metal forming modelling, 4/20–2
Slip rings, copper alloys for, 1/60
Slotting machines, 5/3
SMCs *see* Surface mount components
SMC (sheet-moulding compounds) polyester resin, 2/25–6
SMT *see* Surface mount technology
SMV (server message block) networks, 13/19
SNA (system network architecture), 13/19
Software,
see also Computers
ADAPT, 14/28
Algol, 13/8
APL, 13/8
APT, 14/20, 14/28, 14/56–7
Basic, 11/54, 13/8
C, 11/55
Cobol, 13/7–8
DOS, 14/15–16
EPROM emulation, 11/55
EXAPT, 14/28
Forth, 11/54–5
Fortran, 13/7, 14/16
LISP, 13/24
machine code, 11/55
Pascal, 13/8
PL/1, 13/8
PLCs, 11/60–2
PROLOG, 13/24
redirector, 13/19
UNIAPT, 14/28
UNIX, 14/15, 14/16
VMS, 14/15
Solar heating panels, copper alloys for, 1/62
Soldering,
methods, 9/25–6
PCBs,
conductive belt, 7/12
infra-red soldering, 7/11
laser flow, 7/11
vapour phase reflow, 7/11
wave soldering, 7/10
solders, 1/92, 9/25
Solvent resistance, polymers, 2/20–1
Spark machining *see* Electrical discharge machining
Specifications, quality aspects, 17/7
Specific heat, 1/5
Spectral absoption coefficient, 1/5
Speech recognition, computers, 13/17
Spinning, sheet metal forming, 4/110
Spirit levels, 17/36
Springs,
copper alloys for, 1/61
rubber, 2/38–41
Spring steel *see* Steel
Stability,
amplifiers, 10/24–5
control systems, 11/4–5, 11/9–12
with feeback, 11/11–12
phase advance correction, 11/12
Routh-Hurwitz criterion, 11/12
Stabler's law, rake angles, 5/6
Stainless steel,
AISE/SAE numbering system, 1/11
Allegheny Ludlum A236, 1/38
austenitic, 1/12, 1/37–8, 1/39, 1/40–1
Avesta, 1/38
BSC Esshete, 1250 1/38
Carpenter 20 Cb3, 1/38
cold drawing, 4/97, 4/98
corrosion resistance,
austenitic steels, 1/39
chromium, influence of, 1/39
crevice corrosion, 1/39
ferritic steel, 1/39
galvanic corrosion, 1/39
heat treatment for, 1/39
martensitic steel, 1/39
molybdenum, influence of, 1/39
nickel, influence of, 1/39
pitting, 1/39, 1/42
service temperature, 1/39, 1/42

stress corrosion cracking, 1/39, 1/42
sulphuric acid attack, 1/39
surface conditions, 1/39
testing for, 1/43
duplex, 1/12, 1/38
explosive forging, 4/125, 4/126
explosive welding, 4/122
ferritic, 1/12, 1/36, 1/37
flow forming, 4/51, 4/52
laser-beam machining, 6/7–8
low interstitial grades, 1/36
lubricants when metal forming, 4/32
martensitic, 1/12, 1/36–7
metallurgy, 1/36–8
nickel alloys, 1/38
nickel superalloys, 1/38
Nitronic, 1/38
precipitation hardening, 1/12, 1/38
tubing, 4/48, 4/78, 4/80
weldability,
carbide solution and precipitation, 1/47
ferritic stainless steels, 1/48
knife-line attack, 1/47–8
martensitic stainless steel, 1/48
Schaefler diagram, 1/48
stabilised steels, 1/48
weld decay, 1/47
Standards,
see also British Standards Institution
communication standards for microprocessors, parallel IEEE–488 (Hewlett-Packard), 11/41
RS366 phone dialling, 13/21
RS408 NC interface, 13/21
serial interfaces, RS232C, RS422, RS423, RS449, 11/40–3, 13/21, 13/22
StarLAN networking system, 13/19
Staving, tube extrusion, 4/80
Steatites *see* Ceramics
Steel,
see also Heat treatment; Stainless steel; Steelmaking
for aircraft propulsion turbines, 1/43
AISE/SAE numbering system, 1/11
alloy, cold drawing, 4/97
alloys, function of, 1/30–1
aluminium, influence of, 1/34
antimony, influence of, 1/34
arsenic, influence of, 1/34
austenite, 1/21–5
bainite, 1/24–5
basic oxygen furnace (BOF) steel, 1/27
bismuth, influence of, 1/34
boron, influence of, 1/34
boron intensified steels, 1/11
carbon, influence of, 1/34
carbon-manganese steel forgings, 1/27
carbon steels, 1/10, 1/11
case hardening steels, 1/10
casings, standards for, 1/51
cast,
casting essentials, 1/50
continuous casting, 1/50
creep resistance, 1/50
cementite, 1/21
for chemical and refinary plant, 1/43
chromium, influence of, 1/34
chromium-molybdenum steels, 1/11
chromium steels, 1/11
chromium-vanadium steels, 1/11
classification, BS 970 and BS, 4670 1/32
classification by stength, 1/32–3
coil steel, 1/27
cold reduced strip, 1/27
cold rolled narrow strip, 1/27
copper, influence of, 1/34
creep resisting steels, 1/12
cryogenic applications, 1/45
crystalline structure of, 1/15
electrical, 1/28–9
electric arc furnace, 1/15
electric steels, 1/12
embrittlement, 1/44
eutectoid steel, 1/23
ferrite, 1/21–5
free cutting, 1/31
heat resistance,
applications, 1/43
austenite steels, 1/44
chromium steels, 1/44
embrittlement, 1/43–4
ferritic chromium steels, 1/43
structural stability, 1/43–4
time-dependant deformation, 1/43
valve steels, 1/44
high strength low alloy (HSLA) steels, 1/10, 1/28
Hized plate, 1/26
hot extrusion, 4/78
hot rolled steel, 1/10
hot working, 1/18
hydrogen, influence of, 1/34
hypereutectoid steel, 1/23
hypoeutectoid steel, 1/22
ion implantation, 1/35
iron carbide, 1/21
laser-beam machining, 6/7–8
leaded steels, 1/11, 1/31
lead, influence of, 1/34
low alloy high tensile steels, 1/11
for magnetic cores, 1/28–9
manganese, influence of, 1/34
manganese steels, 1/11
manufacturing processes, 3/4
maraged, 1/12, 1/45–6, 1/47
martensite, 1/23–4
martensitic structures, 1/45
mechanical working, 1/15–21
for metal forming tools,
cold work steel, 4/33
hot-work steel, 4/33
HSS (high speed steel), 4/33
mould steels, 4/33
shock-resisting steel, 4/33
special purpose steels, 4/33
water-hardening steels, 4/33
molybdenum, influence of, 1/34
moybdenum steels, 1/11
nickel-chromium-molybdenum steels, 1/11
nickel-chromium steels, 1/11
nickel, influence of, 1/34
nickel-molybdenum steels, 1/11
nickel steels, 1/11
niobium, influence of, 1/34
nitrogen, influence of, 1/34
for oil pipelines, 1/26
oxygen blown, 1/13
pearlite, 1/22, 1/23
pearlite cementite, 1/23
pearlite ferrite, 1/23
phosphorus, influence of, 1/34
for pressure vessels, 1/26–7
rake angles for machining, 4/5.15
selenium, influence of, 1/34
sheet steel, 1/27
silicon, influence of, 1/34
silicon-manganese steels, 1/11
spring steels, 1/12
austenite steel, 1/50
carbon spring steel, 1/49
cold drawn carbon steel, 1/49
corrosion, 1/50
martensitic steel, 1/50
patenting, 1/49
scragging, 1/50
for steam-power-plant, 1/43
strength classification, 1/32–3
structural plates, 1/26
sulphur, influence of, 1/34
tin, influence of, 1/34
titanium, influence of, 1/34
tool steels, 1/12, 1/48–9
toughness,
Charpy V notch value, 1/45
definition, 1/44
tramp elements, 1/15, 1/45
for tubes, 1/26
tungsten-chromium steels, 1/11
for valves, internal combustion, 1/44
valve steels, 1/12
vanadium, influence of, 1/34
weldability, 1/46–7
weldable structural steel, 1/25
wire, 1/27–8
world output, 3/3–4
Steelmaking,
see also Forging; Heat treatment; Rolling metal; Stainless steel; Steel
Argon oxygen degassing (AOD), 1/13
basic oxygen converter (BOC), 1/13
blast furnace, 1/12
consumable vacuum arc, 1/15
continuous casting, 1/13, 1/14, 1/15
Dortmund Horder (DH) process, 1/13
drawing, 1/21
electric arc heating, 1/13
electric arc process, 3/4
electron beam, 1/15
electroslag remelting (ESR), 1/15, 1/17, 1/21
extrusion, 1/21, 4/78
ingot casting, 1/13, 1/15
in-works analysis, 1/13
oxygen argon lancing, 1/13
oxygen converter process, 3/4
Ruhr-Stahl Heraus (RH) process, 1/13
tramp elements, 1/15
vacuum arc deoxidation (VAD), 1/13, 1/16
vacuum carbon deoxidation, 1/13
vacuum degassing, 1/13
vacuum high frequency furnace, 1/15
Steel rules, 17/34
Stepper motors,
applications and principle, 10/57–8
CNC machines, 14/31
hybrid, 10/58
microprocessor speed control, 11/58
permanent magnet, 10/58
variable reluctance, 10/58
Stiefel system, metal rolling, 4/43
Stock *see* Inventory
Strain hardening, from rolling, 4/38
Strain measurement *see* Transducers
Strain redundancy factor, plasticity, 4/18
Stress-strain relationships,
see also Plastic flow of metal; Plasticity
bulk modulus, 1/3
elastic or Young's modulus, 1/3
fatigue endurance, 1/3
flexural strength, 1/3
limit of proportionality, 1/3
Poisson's ratio, 1/3, 4/19
shear modulus, 1/3
stress redundancy factor, plasticity, 4/18
stress to a certain creep strain, 1/4

Stress-strain relationships, *continued*
stress to rupture, 1/4
tensile ductility, 1/3
testing methods, metal, 4/11
ultimate tensile stress, 1/3
Stress technique, metal forming modelling, 4/19–20
Strip, rolling *see* Rolling metal
Structural sections, rolling, 4/42
Stylus measurements, 17/39–40
Styrene-butadiene rubber, 2/13–14
Styrene-butadiene-styrene triblock rubbers, 2/15
Superabrasives, 8/6, 8/7–8
Superfinishing,
cycle, 8/25
machining mechanisms, 8/23–5
machining set-up, 8/23–5
Superheated steam fittings, copper alloys for, 1/62
Superplastic metal forming, 4/126–7
Superposition theorem, 10/4
Supplier assessment, 15/54
Surface finish,
electrical discharge machining, 6/3–4
electrochemical cleaning, 6/4
machining, 5/17
ultrasonic machining, 6/6
Surface mount components (SMCs),
active,
cylindrical diode packages, 7/4
leadless ceramic chip carriers (LCCC), 7/4
plastic leaded chip carriers (PLCC), 7/4
quad flat packs, 7/4
small outline integrated circuit (SOIC), 7/4
small outline transistors (SOT), 7/4
board attachment, 7/9
packaging,
bulk feed, 7/6
matrix trays, 7/6
sprocket driven tape, 7/6
sticks, 7/6
waffle packs, 7/6
passive,
chip capacitors, 7/5–6
chip resistors, 7/5–6
solder pastes for, 7/10
Surface mount technology (SMT), 7/4
introduction of, 7/13
Surface plates, 17/35
Swarf disposal, 5/5, 5/14
Switch blades, copper alloys for, 1/60
Switches,
limit, 12/13–14
reed, 12/14
Synchronous motors,
microprocessor speed control, 11/58
operation, 10/53
torque,
running, 10/53
starting, 10/53–4
System application architecture (SAA), 13/19
System architecture, microprocessors, 11/37
System network architecture (SNA), 13/19
Systems theory, production management *see* Management

Tachogenerators, 11/26
Tantalum,
characteristics, 1/93
tantalum carbide, 1/107
Tape,
magnetic, 14/51
paper, 14/50
streamers, 14/15, 14/22
Tape punches,
for CADCAM, 14/15
for CAM, 14/22
Taper limit gauges, 17/37
Taps, copper alloys for, 1/62
Taylor mill, 4/37
Taylor's theorem of gauging, 17/36–7
Telenet network, 13/22
Temperature measurement,
platinum-resistance thermometer, 11/31
thermisters, 11/31
thermocouples, 11/31
Temperatures, metal forming, 4/4–5
Tensile ductility, 1/3
Tensile testing, 1/3
Terminals, CADCAM, 14/15
Terne plate, 1/92
Terotechnology,
see also Maintenance
asset management, 18/3
maintenance and, 18/4
practice, 18/3
Testing,
eddy-current,
fixed head technique, 17/55–6
search head technique, 17/56
magnetic particle, 17/52, 17/54
PCBs, 7/12
penetrant, 17/54–5
Therbligs, motion study, 15/18
Thermal properties,
specific heat, 1/5
thermal conductivity, 1/5
thermal diffusivity, 1/5
thermal expansion, 1/5
Thermal stability, plastics, 2/16–17
Thermoforming,
plastics and rubbers,
compressed air, 2/36
mechanical shaping, 2/35
vacuum forming, 2/36
Thermoplastic rubbers, 2/5, 2/15
Thermoplastics, 2/3
amorphous, 2/3–4
biaxial orientation, 2/4
complex orientation, 2/4
glass transition temperature, 2/4
monoaxial orientation, 2/4
triaxial orientation, 2/4
crystalline, 2/4
melting point, 2/4
engineering types of, 2/24–5
major tonnage types of, 2/22–4
Thermosetting plastics, 2/3, 2/25–6
fibre-reinforced, 2/36–7
Thermosetting polyester resins, 2/25–6
Thevenin's theorem, 10/4
Threads,
anaerobic cement locking, 9/6–7
ISO metric screw thread system, 9/4
liquid adhesive locking, 9/6–7
measurement, 17/43–6
screw thread gauges, 17/37
sealing, 9/32
threaded inserts for plastics, 9/7
threadlocking, 9/6–7
Unified Coarse (UNC), 9/4
Unified Fine (UNF), 9/4
Three-phase,
alternators, 10/53
generation, 10/51, 10/52
power in, 10/52–3
star and delta connection, 10/51–2
voltage and current relationships, 10/52
Thyristors, 10/22–3, 10/29
see also Silicon controlled rectifier
phase angle control by, 11/45
Time constants, electrical circuits, 10/12
Time domain analysis,
damping, 11/6–8
expontial lag, 11/6
first order systems, 11/6–7
logarithmic decrement, 11/8
second order systems, 11/6–8
Timers, logic systems, 10/42
Time-temperature transformation (T-T-T), steel, 1/24
Tin/tin alloys,
applications, 1/92
as bearing material, 1/92
forging, 4/63
solder, 1/92
standards, 1/92
terne plate, 1/92
tin plate, 1/92
Titanates, 1/111
Titania ceramics, 1/111
Titanium carbide/nitride tool coatings, 5/11
Titanium/titanium alloys,
alloy structure, 1/70–2
applications, 1/70, 1/75
BS specifications, 1/74
centrifugal casting, 1/75
chemical composition, 1/74
creep properties, 1/75
fatigue crack growth, 1/72
forging, 1/75
fracture toughness, 1/72
general characteristics, 1/69
heat treatment, 1/72
hot extrusion, 4/78
IMI numbers, 1/74
laser-beam machining, 6/7
nomenclature, 1/74
phase diagrams, 1/70
stabilising and strengthening elements, 1/70–2
welding, 1/75
Token passing, network, 13/19
Token ring, network protocol, 13/19
Tooling,
CNC turning centres, 14/42–3
management and supply,
automatic distribution, 14/65
cell centred, 14/64
machine centred, 14/64
system centred, 14/65
requirement planning, 15/17–18
Tools/tool design,
cutting tools,
see also Machining
carbide tools, 5/11, 5/15
cast non-ferrous alloys, 5/10
catastrophic failure, 5/19
ceramic, sintered, 5/11
chatter, 5/11, 5/25
chip formation, 5/11–14
coolants, 5/19
cubic boron nitride, 5/11
cutting parameters, 5/11–14
diamond, 5/11
flank wear, 5/19
fluids for, 5/13–14
heat generated, 5/13
high speed steel, 5/10
life assessment, 5/17–20
lubrication, 5/13–14

materials for, 5/10–11
milling forces and power, 5/15–17
multipoint cutters, 5/25
orthogonal cutting, 5/6
plastic deformation, 5/18–19
polycrystalline synthetic diamond, 5/11
rake, 5/6–10, 5/13, 5/14, 5/15
silicon nitride, 5/11
sintered carbides, 5/10–11
Stabler's law, 5/6
swarf disposal, 5/14
thermal shock in, 5/19
titanium carbide/nitride coatings, 5/11
tool changing times, 5/22
tool geometry, 5/6–10
turning forces and power, 5/14–15
vibration in, 5/11
wear, cratering, 5/18
wear, crumbling, 5/18
electrical discharge machining, 6/3–4
metal forming,
cemented carbide composites for, 4/33
ceramic materials for, 4/33
CMSR (constant mean strain rate), 4/28, 4/30
CRHS (constant rate of homogeneous strain), 4/28, 4/29–30
flow-line concept, 4/30–1
forging, 4/27, 4/70–2
fracture and wear in, 4/32
lateral shape factor, 4/70
lubricant contamination, 4/31
lubrication, 4/31–2
materials for, 4/32–3
metallic film lubricants, 4/31
non-metallic film lubricants, 4/31–2
residual stresses in, 4/33–4
rotary piercing, 4/28
steels for, 4/33
surface treatment of, 4/33
surface treatments for, 4/33
titanium carbide composites for, 4/33
Ugine-Sejournet process, 4/32
non sparking, copper alloys for, 1/61
steels for,
carbon steels, 1/48
molybdenum steels, 1/48
tungsten steels, 1/48
tube rolling, 4/48–51
Topology, network, 13/19
Toroid, magnetic field of, 10/7
Touch screens, computers, 13/13–14
Toughness, plastics, 2/18–19
Tracker balls, CADCAM, 14/26
Trades Description Act, 17/6
Tramp elements, steel, 1/15, 1/45
Transducers,
see also Sensors; Temperature measurement
absolute shaft encoder, 11/23–5, 12/19–20
acceleration, 11/28–9
accelerometers, 11/29
amplifiers for, 11/33–4
analogue, 11/22
analogue position, 11/22
analogue speed, 11/22
angular position, 11/23–5
capacitive sensors, 12/14–15
digital, 11/23
filters for, 11/34–6
flowmeters,
Laser-Doppler, 11/30
pressure differential, 11/29, 11/30
rotary, 11/29–30
ultrasonic beat frequency, 11/30–1
vortex shedding, 11/30
fluid level sensing, 11/31–2
force, 11/28–9
frequency-to-voltage converter, 11/25
Hall effect sensors, 11/29
incremental linear encoder, 11/25
incremental shaft encoder, 11/23–5, 12/19
inductive sensors, 12/14
inductosyns, 11/25, 12/17, 14/32
linear, CNC machines, 14/32
linear potential divider, 11/25, 12/16
linear variable displacement transformer (LVDT), 11/25
magnetic pick-up speed sensor, 11/26
microswitches, 12/14
optical linear encoders, 12/17–20
optical shaft encoders, 12/17–20
photosensors,
cadmium sulphide, 12/15
photosensitive semiconductors, 12/15
photovoltaic, 12/15
piezoelectric effect, 11/28–9
pneumatic sensors,
back pressure, 12/15
interruptible-jet, 12/16
position switches, 12/13–14
potentiometers, 12/16
pressure, 11/28–9
proximity sensing/sensors, 11/29, 12/14–16
reed switches and magnets, 12/14
resolvers, 12/17
rotary, 12/16
CNC machines, 14/32
slotted opto-switch, 11/26
strain measurement,
bonded foil gauges, 11/27–8
electrical bridges for, 11/26–8
temperature compensation for, 11/28
tachogenerator, 11/26
Transfer moulding, plastics and rubbers, 2/28–9
Transformers,
autotransformers, 10/61–2
efficiency, 10/60–1
losses,
copper, 10/60–1
iron, 10/61
principle of operation, 10/59–60
referred values, 10/60
three-phase, 10/61–2
voltage equation, 10/60
voltage regulation, 10/61
Transistor,
circuit characteristics, 10/21
common base characteristics, 10/19–20
common emmitter characteristics, 10/19–21
field effect (FET), 10/21–2
junction field effect (JFET), 10/22
MOSFET, 10/22
principle, 10/18–19
Transistor-transistor logic (TTL) gates, 10/30–1
Transport layer, OSI, 13/20
Tresca criterion of plastic flow yield, 4/17–18
Triac, 10/23
Truth tables, 10/30, 10/32
TTL (transistor-transistor logic) gates, 10/30–1
Tube drawing, strain and strain-rate relationships, 4/16–17
Tube flanging, 4/110
Tube flow forming, 4/51–3
floturning, 4/51, 4/53
par-forming, 4/52
Tube rolling, 4/6–10, 4/42–8
Assel elongating, 4/42, 4/45–6, 4/48, 4/49
Calmes elongator, 4/47
cavitation, 4/48–50
cross rolling, 4/42, 4/45–8
Diescher elongating, 4/42, 4/48
ovality, 4/48
piercing, 4/46–8, 4/48
plug mill system, 4/43
push-bench process, 4/45
reducing mills, 4/43–5
reducing sinking mills, 4/44
stainless steel, 4/48
Stiefel system, 4/43
theory, 4/55
tooling for, 4/48–51
triangulation, 4/46–7, 4/48–51
wall lamination, 4/48
Tube rotary forging, 4/66–9
Tubes, copper alloys for, 1/62
Tungsten,
characteristics and applications, 1/93
tungsten carbide, 1/107
Turning centres, CNC,
applications, 14/39–41
axes, 14/42
in flexible manufacturing, 14/63
program verification, 14/43
tooling, 14/42–3
work handling, 14/43–4
Tyre rubbers, 2/13–14

Ugine-Sejournet lubrication, 4/32, 4/79
Ultrasonic machining,
abrasive materials used, 6/6
applications, 6/5
limitations, 6/5
principle, 6/5–6
surface finish, 6/6
Ultrasonics,
beam spread, 17/51
energy propagation, 17/50
flaw detection,
reflection technique, 17/52, 17/53
transmission technique, 17/52, 17/53
Fresnel zone, 17/51
generation, 17/51–2
Lamb waves, 17/50
lenses, 17/51
longitudinal waves, 17/50, 17/51
piezoelectric effect, 17/52
probe types, 17/52
refraction, 17/50–1
shear waves, 17/50, 17/51
surface waves, 17/50, 17/51
wave motion, 17/50
wave trains, 17/50
Ultrasonic vibratory tube and wire drawing, 4/97
Underwriter Laboratory, UL 94 Vertical Burning Test, 2/17
Unified Coarse (UNC) threads, 9/4
Unified Fine (UNF) threads, 9/4
Upper-bound method, modelling metal formation, 4/22
Upsetting, forging, 4/57–9
UPVC (unplasticised PVC), 2/22
Urea-formaldehyde (UF) plastics, 2/25
U-rings, rubber, 2/41–2
Usage rate, equipment, 18/6

Vacuum Steelmaking processes *see* Steelmaking
Valve bodies, copper alloys for, 1/61
Valves, control *see* Control valves

Valves, internal combustion, steels for, 1/44
Valve steels *see* Steel
Valves, water, copper alloys for, 1/62
Vector display monitors, CADCAM, 14/25–6
Vendor assessment,
  height gauges, 17/34
  purpose and method, 17/13, 17/15
  vetted suppliers, 17/15
Vernier calipers, 17/34
Vibration, in machine tools, 5/11
Vibration isolation, with rubber, 2/41–2
Vicat (plastic) softening point, 2/16
Vickers hardness, 1/3–4
Vision systems,
  applications,
    decision making, 14/86
    identification, 14/85
    inspection, 14/85–6
  components, 14/85
  operation, 14/85
  types, 14/85
Visioplasticity technique, modelling metal formation, 4/22–3
Vitrification, ceramics, 1/115
VLSI (very large scale integration) circuits, 11/37
Voice input, computer, 13/16–18, 14/87
Volts, 10/4
Volume resistivity, 1/5
Vulcanisation, 2/5

Wall ties, copper alloys for, 1/62
WANs (wide area networks), 13/22
Washers, 9/7, 9/8
Water cylinders, copper alloys for, 1/62
Water-jet machining, 6/5
Waves, sound *see* Ultrasonics
WCM (world class manufacturing), 16/7
Weldable structural steels, 1/25
Welding,
  see also Weld radiographs
  dies, copper alloys for, 1/62
  economics of, 9/26–7
  electrodes, copper alloys for, 1/62
  electron beam,
    high vacuum, 9/22
    medium vacuum, 9/22
    out of vacuum, 9/22
  electroslag,
    fluxes, 9/17
    mechanical properties, 9/17
    process, 9/16–17
    process economics, 9/17
  explosive, 4/105, 4/121
  ferritic stainless steels, 1/48
  flash, 9/24–5
  flux cored arc,
    applications, 9/19
    principle, 9/19
    tubular cored electrodes, 9/19
  friction, 9/25
  gas,
    applications, 9/21
    equipment, 9/21
    filler metal and fluxes, 9/21
    principle, 9/21
  gas shielded metal arc, 9/17
  gravity, 9/14–15
  heat-affected zone (HAZ), 9/14
  intergranular corrosion, 1/47
  joint types,
    butt, 9/11
    lap, 9/11
  knife line attack, 1/47
  laser,
    carbon dioxide, 9/23
    pulsed, 9/22–3
    yttrium-aluminium-garnet (YAG), 9/22–3
  manual metal arc,
    applications, 9/14
    BB (basic, high efficiency) flux coverings, 9/14
    C (cellulosic) flux coverings, 9/14
    electrodes, 9/13, 9/14
    MR (rutile, heavy coated) flux coverings, 9/13–14
    power sources, 9/14
    process, 9/12–13
    R (rutile) flux coverings, 9/13
  martensitic stainless steels, 1/48
  MIG/MAG (metal inert/active gas),
    applications, 9/19
    filler wires, 9/17–18
    globular metal transfer, 9/18
    power sources, 9/18–19
    principle, 9/17
    pulsed metal transfer, 9/18
    short circuiting metal transfer, 9/18
    spray metal metal transfer, 9/18
  open arc automatic, 9/15
  plasma arc, 6/8
    keyhole technique, 9/20
    principle, 9/20
  precipitation, 1/47
  processes, classified, 9/11, 9/12
  productivity, 9/26–7
  projection, 9/23–4
  resistance, 9/23–4
  resistance butt, 9/24
  Schaefler diagram, 1/48
  seam, 9/23
  spot, 9/23, 14/81
  steels for,
    non-stainless, 1/46–7
    stainless, 1/47–8
  submerged arc,
    acid fluxes, 9/16
    agglomerated fluxes, 9/16
    applications, 9/16
    basic fluxes, 9/16
    electrode wires, 9/15–16
    fused fluxes, 9/16
    neutral fluxes, 9/16
    power sources, 9/16
    principle, 9/15
  technology background, 9/11
  TIG (tungsten) inert gas,
    applications, 9/20
    electrodes, 9/20
    filler rods, 9/20
    power sources, 9/20
    principle, 9/19–20
  titanium, 1/75
  T joints, 9/11
    corner joints, 9/11
  weld decay, 1/47
Weld radiographs,
  cracks, 17/49
  incomplete penetration, 17/49
  incomplete side wall fusion, 17/49
  piping, 17/49
  porosity, 17/49
  slag inclusions, 17/48
  undercutting, 17/49
Window frames, copper alloys for, 1/62
Wire, steel, 1/27–8
Wire drawing *see* Cold drawing
Wood, water-jet machining, 6/5
Wood composites,
  compressed impregnated, 1/108–9
  laminate strengths, 1/108–9
  plywood, 1/107
  veneers, 1/107–8
Work change stations, flexible manufacturing, 14/63
Workstations,
  CADCAM, 14/15
  CAM, 14/22
Work study,
  Gilbreth's work, 15/18
  motion study, 15/18
  operation analysis sheet, 15/19
  purpose, 15/18
World class manufacturing (WCM), 16/7
Wrought iron, 1/107

X.25 packet switching network, 13/20
XNS (Xerox network services), 13/20
XPS (expanded polystyrene), 2/22
X-rays,
  X radiation, 17/46, 17/48
  X-ray equipment, 17/46–7

Yielding, criteria of in plastic flow, 4/17–18
Young's modulus, 1/3, 2/7–8

Zalutite, 1/92
Zener diode, 10/18
Zinc/zinc alloys,
  applications, 1/87
  casting alloys, 1/90
  characteristics, 1/87
  corrosion protection coatings, 1/90–2
  electrogalvanising, 1/92
  hot dip galvanising, 1/90
  properties, 1/91
  standards, 1/87, 1/90
  wrought alloys, 1/90
  Zalutite, 1/92
Zirconia ceramics *see* Ceramics
Zirconium, nuclear application, 1/94
z-transforms, 11/21